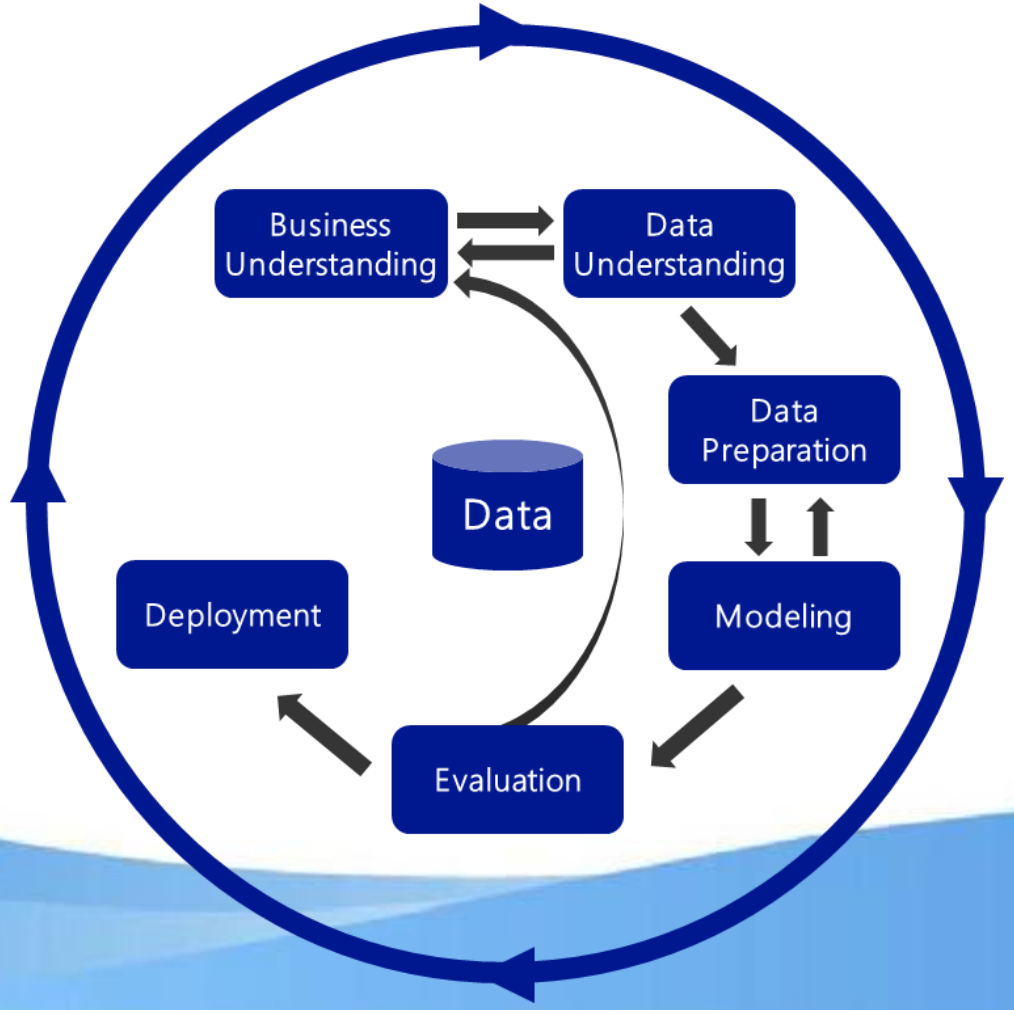
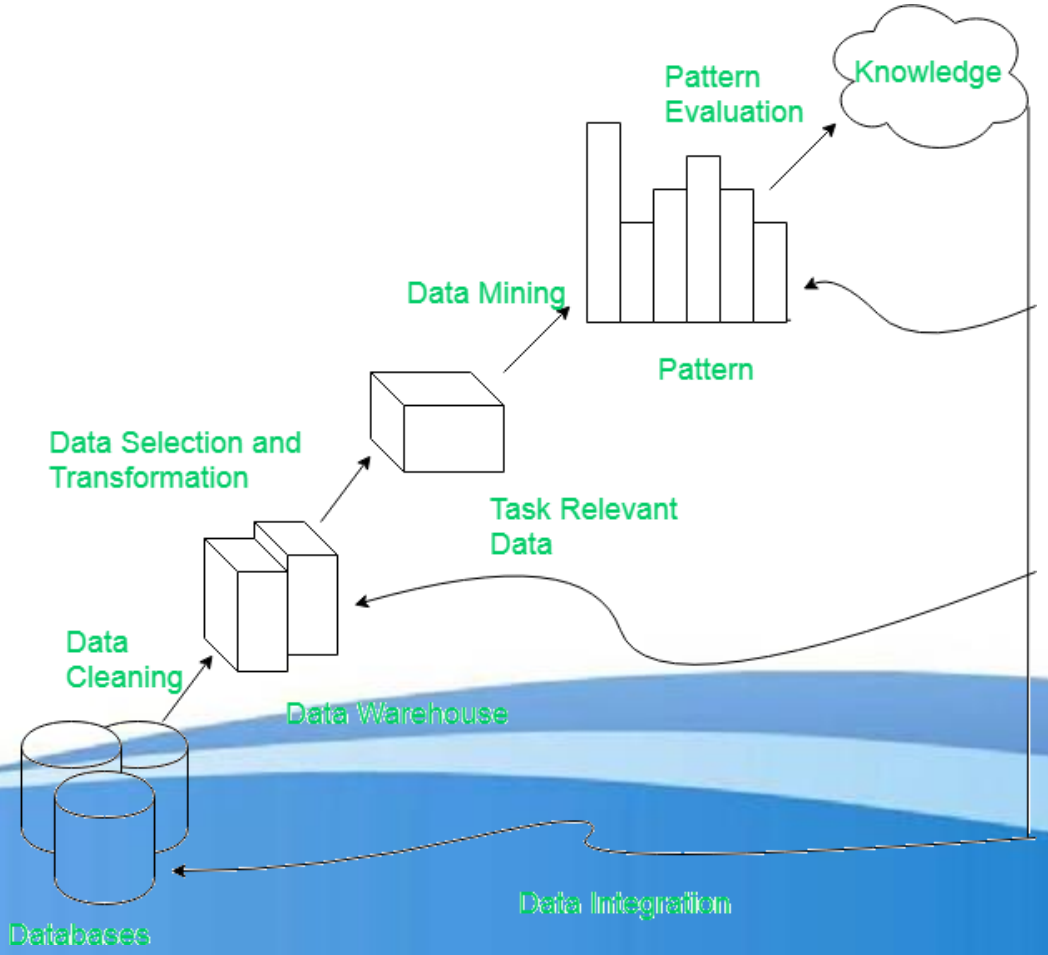
데이터 분석 표준 프로세스 (CRISP-DM)

데이터 마이닝 표준 프로세스

Cross-Industry Standard Process for Data Mining



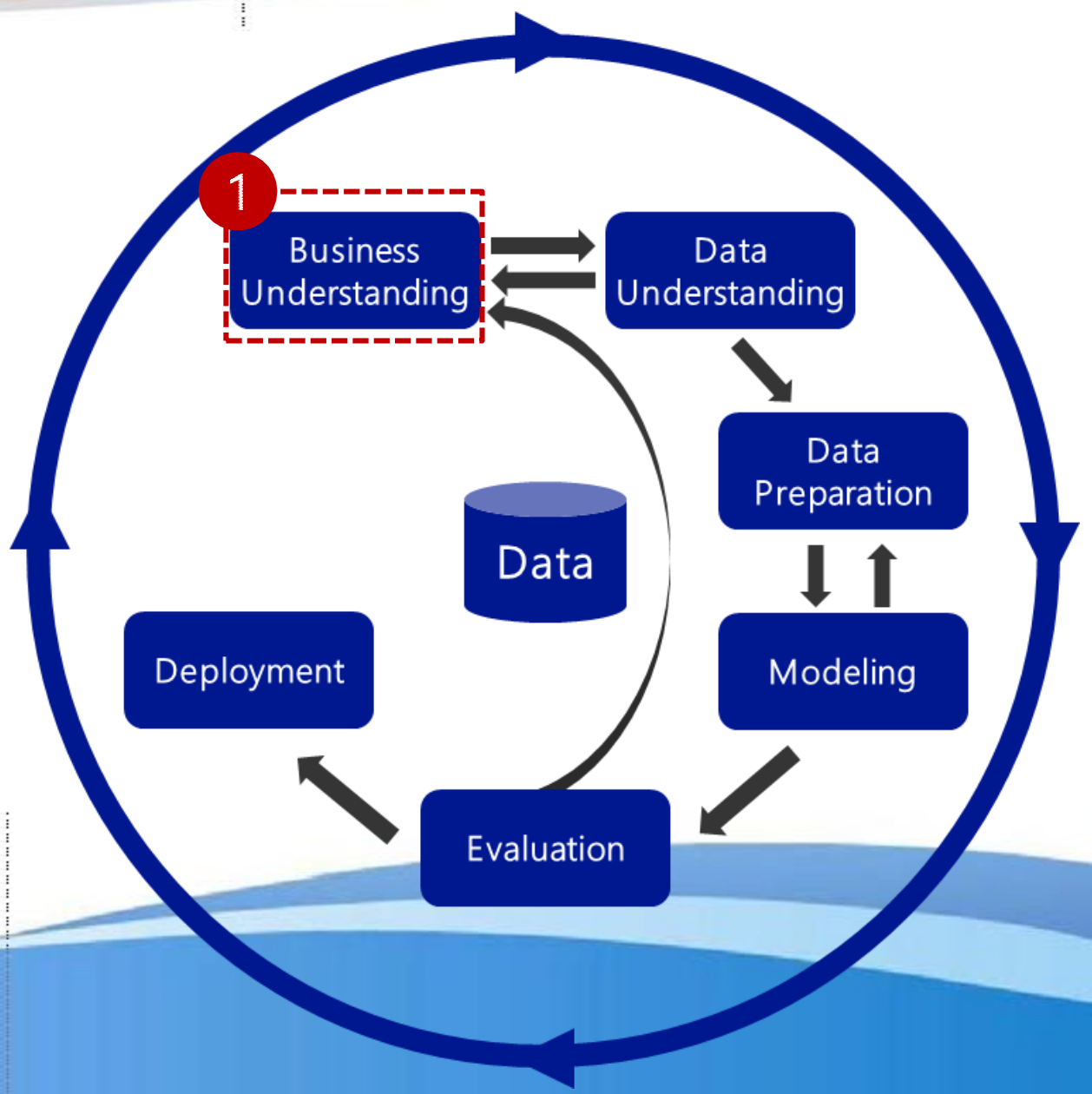
Knowledge Data Discovery



CRISP-DM

- 프로세스를 한번 거쳤음에도 문제가 해결되지 않을 수 있다
  ➔ 그렇다고 실패가 아님!

- 한번에 해결책을 찾지 못해도 데이터를 더 잘 이해하게 되는 계기가 됨
  ➔ 두번째 수행할 때는 더 많은 정보를 갖고 시작할 수 있음!



①Business Understanding

- 개요
  - 잘 정의된 명확한 데이터분석 문제로 시작하는 프로젝트는 거의 없음.
  - 문제를 파악해 가는 과정을 반복하면서 문제를 재정의하고 해결책을 정의하게 됨.

- 수행되는 내용
  - 비즈니스 목표 검토
  - 데이터 분석 목표 수립
  - (초기)가설 수립

①Business Understanding

- 비즈니스 목표에서 데이터 분석 목표로

비즈니스 관점	목표	**2020**년 은행 대출 부서의 수익 **1000**억 달성
	방법	✓ 신용도 높은 사람의 대출 신청 승인 ✓ 신용도 낮은 사람의 대출 신청 거절
데이터 분석 관점	문제정의	대출 신청자들의 신용도를 예측할 수 있을까**?**
	목표	어느 정도 정확도로 예측할 수 있다면**,** ▪ 비즈니스 목표를 달성 할 수 있을까**?** ▪ **2**년 이내 프로젝트 투자에 대한 **BEP**에도달할 수 있을까**?**
	분석	▪ 분류문제 ▪ 신용도에 영향을 미치는 요인은 무엇일까**?**

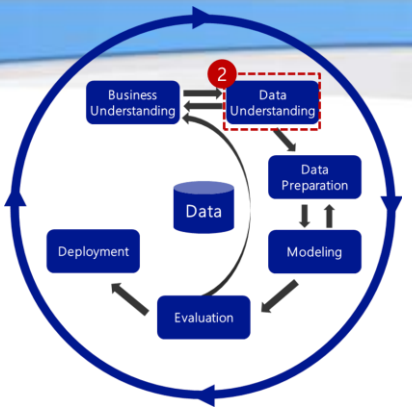
①Business Understanding

- (초기)가설 수립

신용도에 영향을 미치는 요인은 무엇일까?

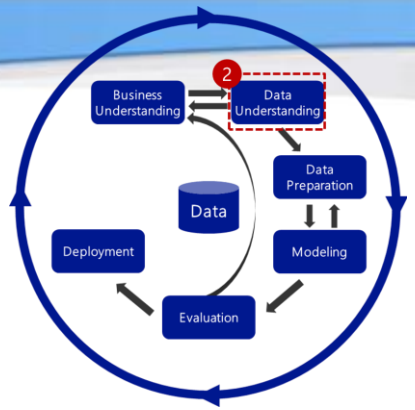
- 다양한 직무에 있는 사람들의 의견을 수렴할 필요가 있음.
- 데이터의 존재여부를 고려하지 말고 가설 도출.
- 초기 가설 수립 이후 데이터 탐색을 통해 가설을 구체화

②Data Understanding



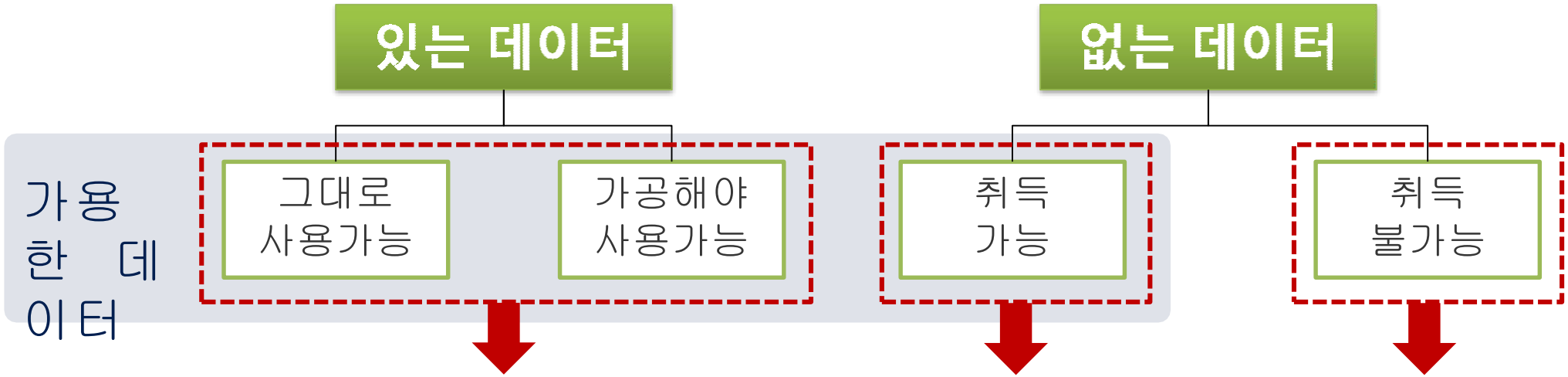
- 개요
  - 데이터 : 문제의 해결책을 만드는 데 사용할 원자재
  - 문제에 정확히 부합하는 데이터가 있는 경우는 거의 없음.
  - 데이터에 따라 데이터 취득 및 유지 비용이 다름.

- 수행되는 내용
  - 데이터 원본 식별 및 취득
  - 데이터 탐색 : EDA, CDA



②Data Understanding

- 데이터 원본 식별 및 취득
  - (초기)가설에서 도출된 데이터의 원본을 확인



Data Pipeline Data Environment 검토 및 과제 수행

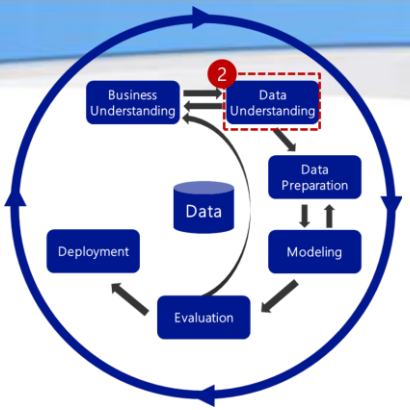
취득 비용 산정 과제 수행

정보의 의미 정의

❼ 정보 분할

❼ 최대한 가용한 ▩ 영역으로 만들기

②Data Understanding

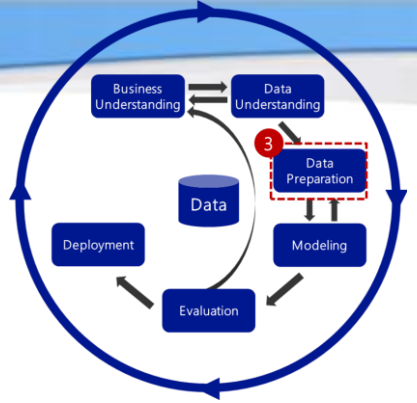


- 데이터 탐색 : EDA, CDA
  - 데이터를 탐색하는 두 가지 방법

데이터 통계량	데이터 시각화
분할표(**Contingency Table**) **MIN, MAX, SUM, MEAN, Quartile …**	**Histogram, Boxplot, Density plot** **Barplot, Pie chart** **Scatterplot …**

  - EDA (Exploratory Data Analysis)
    - 개별 데이터의 분포, 가설이 맞는지 파악
    - NA, 이상치 파악
  - CDA (Confirmatory Data Analysis)
    - 탐색으로 파악하기 애매한 정보는 통계적 분석 도구(가설 검정) 사용

③ Data Preparation



- 개요
  - 데이터 분석을 위해 특정 조건에 맞는 데이터 유형과 구조가 있음
  - 더 좋은 결과를 얻을 수 있도록 데이터의 형태를 조작하고 변환하는 과정 필요.

- 수행되는 내용
  - 데이터 정제
  - 추가 변수(Feature Engineering)

- 결과물 : 하나의 잘 정리/정제된 데이터프레임(테이블)

분석할 수 있는 데이터?

✓연속형
- 숫자
- 날짜

✓예
- 주문일
- 판매량
- 금액
- 나이

✓ 범주형
- 순서형, 명목형, 이항형

✓ 예
- 상품카테고리
- 성별
- 고객
- 지역
- 연령대

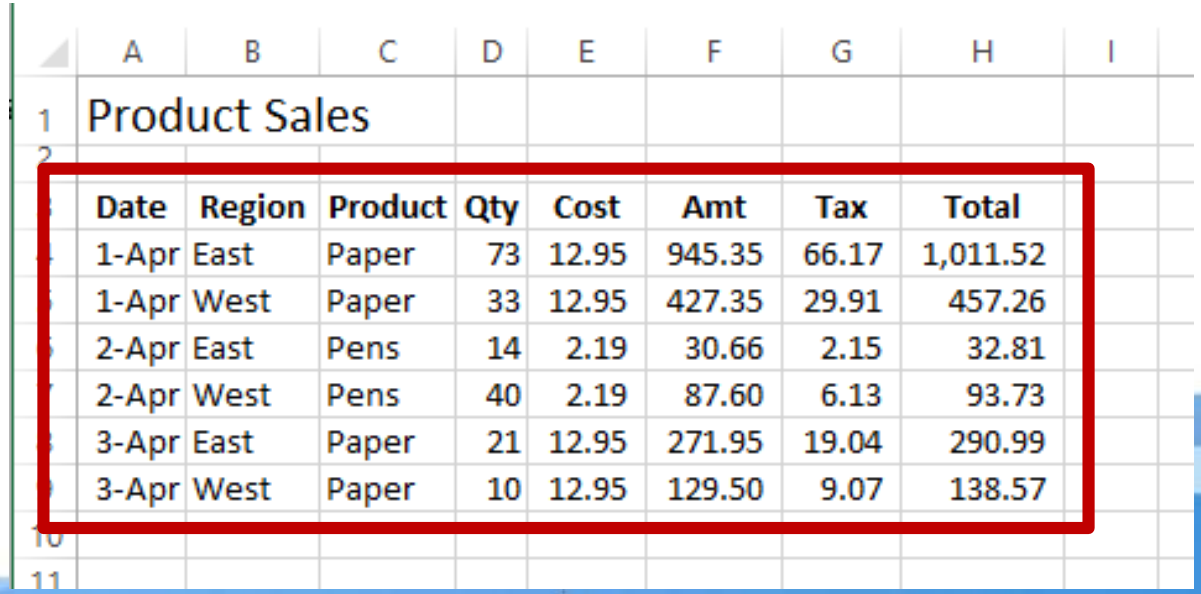
분석을 위한 데이터 구조

✓ Table 형태

① 데이터가 아래로 쌓이는 구조

② 열(column) 성격에 맞는 데이터가 입력되어야 함

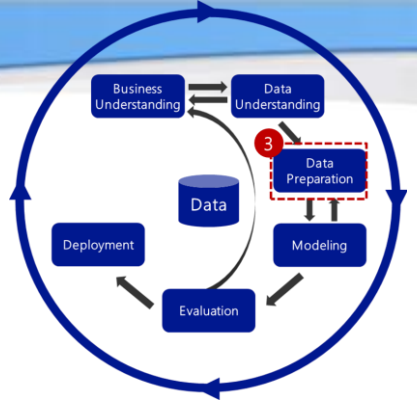
③ 열(column) 마다 데이터 형식이 있음



Product Sales

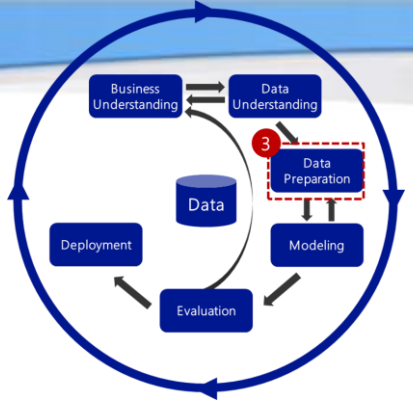
Date	Region	Product	Qty	Cost	Amt	Tax	Total
1-Apr	East	Paper	73	12.95	945.35	66.17	1,011.52
1-Apr	West	Paper	33	12.95	427.35	29.91	457.26
2-Apr	East	Pens	14	2.19	30.66	2.15	32.81
2-Apr	West	Pens	40	2.19	87.60	6.13	93.73
3-Apr	East	Paper	21	12.95	271.95	19.04	290.99
3-Apr	West	Paper	10	12.95	129.50	9.07	138.57

③ Data Preparation



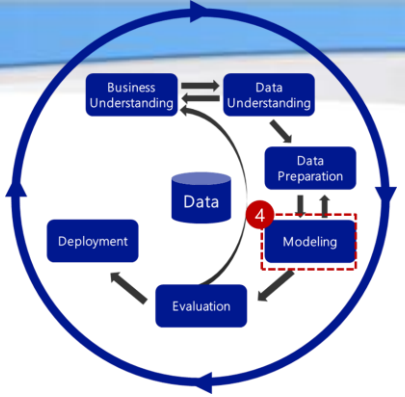
- 데이터 정제
  - 잘못된 데이터 정제
  - 결측치(NA) 식별 및 조치
    - 중요한 요인에 결측치가 존재한다면 반드시 조치해야 한다.
    - 예 : 옷을 추천하는데, 고객의 나이나 성별에 결측치가 존재한다면, 옷을 추천하기 곤란.
  - 이상치 식별 및 조치
    - 잘못된 값
    - 값 자체는 정상이나 다른 값들의 분포에 비해 치우친 값
    - 이러한 값은 데이터 분석 시 잘못된 결과를 얻게 하는 원인이 됩니다.

③ Data Preparation



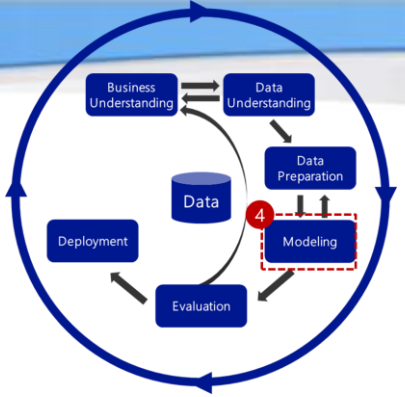
- 추가변수(Feature Engineering)
  - 기존에 저장된 데이터를 그대로 사용해서는 제대로 된 예측 결과를 얻기 어렵다.
  - 데이터베이스에 데이터를 저장하는 방식
    - 트랜잭션 발생 순으로 저장 ➔ 저장된 데이터 자체가 비즈니스의 Insight가 되지 못함.
  - 비즈니스의 경험 + 데이터 분석을 통해 인사이트를 발견하고, 이를 담아내는 정보가 필요
  - 사례
    - 페이스북 고객 중 가입 후 10일 이내 7명의 친구를 사귄 사람은 그렇지 않은 사람보다 잔존율이 훨씬 높다!
    - 음주 습관에 대한 분석 : age 변수를 이용해서 age >= 20 ➔ 음주가능연령
    - 아파트가격 석 : 방 수 >=4 & 화장실수 >=2 ➔ Premium

④ Modeling



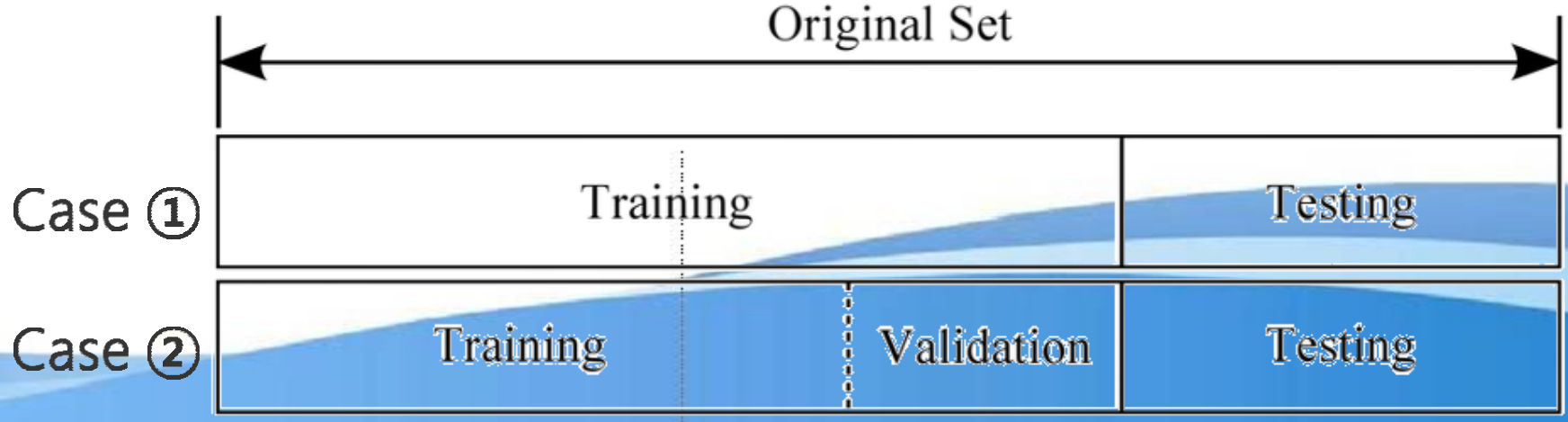
- 개요
  - 중요 변수들을 선택하고, 적절한 알고리즘을 적용하여 예측 모델을 생성
  - 생성된 모델을 평가

- 수행되는 내용
  - 데이터셋 분리
  - 중요 변수 선정
  - 머신러닝 알고리즘 적용하여 모델 생성
  - 모델 테스트

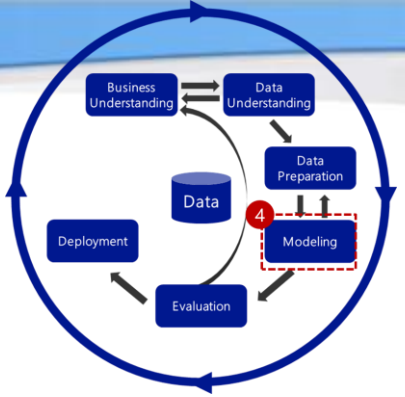


④ Modeling

- 데이터 셋 분리
  - Case ① : 학습할 때
    - Train Set : 알고리즘을 이용해서 모델을 생성
    - Test Set : 모델 성능 검증
  - Case ② : 실전에서 주로 사용
    - Validation Set : 모델 성능 검증
    - Test Set : 모델 최종 평가



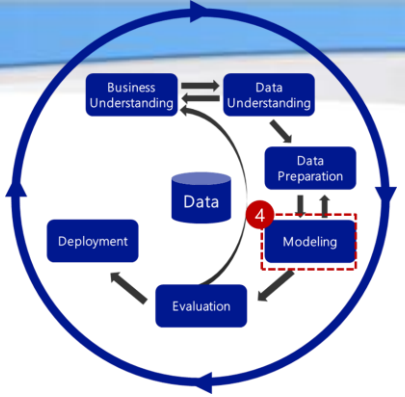
④ Modeling



- 머신러닝 알고리즘

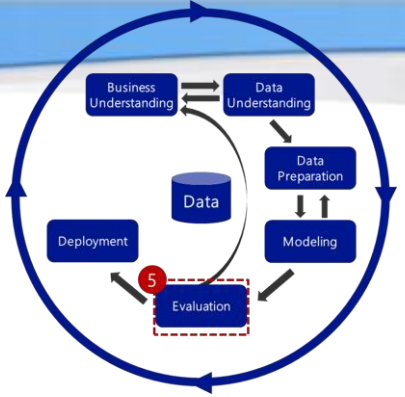
Supervised Learning	Unsupervised Learning
지도학습, 감독학습	비지도학습, 비 감독학습, 자율학습
Label이 있다.	Label이 없다.
Regression, Logistic Regression , SVM, KNN, Decision Tree , Neural Net, Random Forest 등	Clustering : K-Means, DBSCAN 등

④ Modeling



- 모델 생성

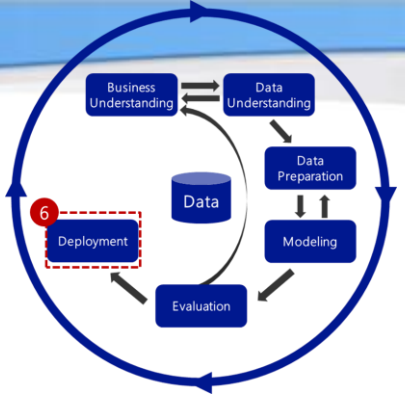
⑤ Evaluation



- 개요
  - 모델에 대한 데이터 분석 목표와 비즈니스 목표달성에 대한 평가
  - 모델과 데이터에서 추출한 패턴이 진정한 규칙성을 갖고 있는지, 단지 특정 예제 데이터에서만 볼 수 있는 특이한 성질은 아닌지 확인
  - 비즈니스 목표에 부합되는지 보장

- 수행되는 내용
  - 모델에 대한 최종평가 : Test Set 이용
  - 비즈니스 기대가치 평가

⑥ Deployment



- 개요
  - 프로젝트 결과물 최종 확정: 프로덕션 환경의 파이프라인, 모델 및 배포가 고객 목표를 충족하는지 확인
  - 운영시스템에서 품질(성능 목표) 유지 기준을 정하고, 모니터링 계획을 수립

- 수행되는 내용
  - 시스템 유효성 검사: 배포된 모델과 이 고객 요구 사항을 충족 하는지 확인
  - 프로젝트 이전 : 운영환경으로 배포