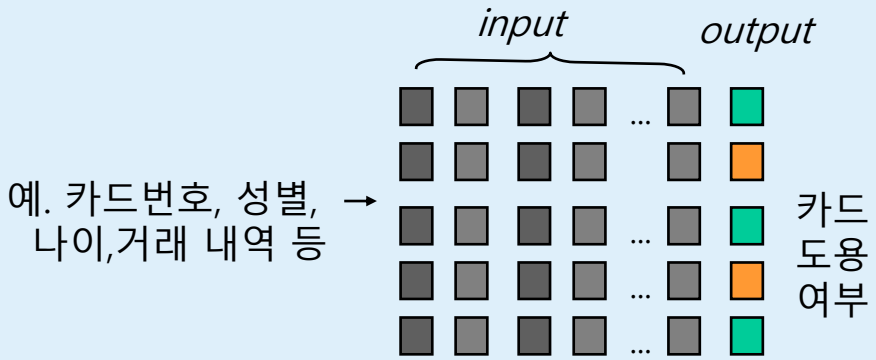
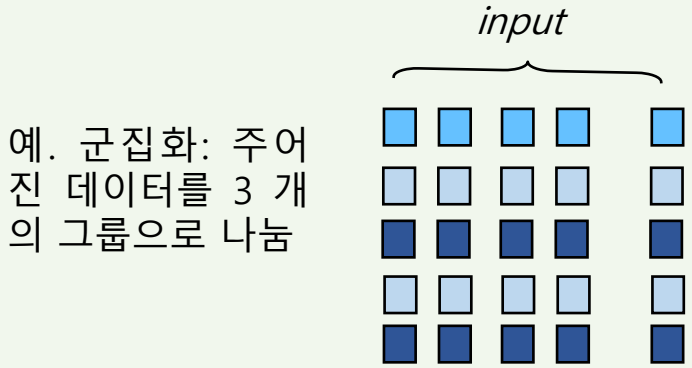
Ⅲ. 데이터 마이닝 유형

1 지도학습과 자율학습

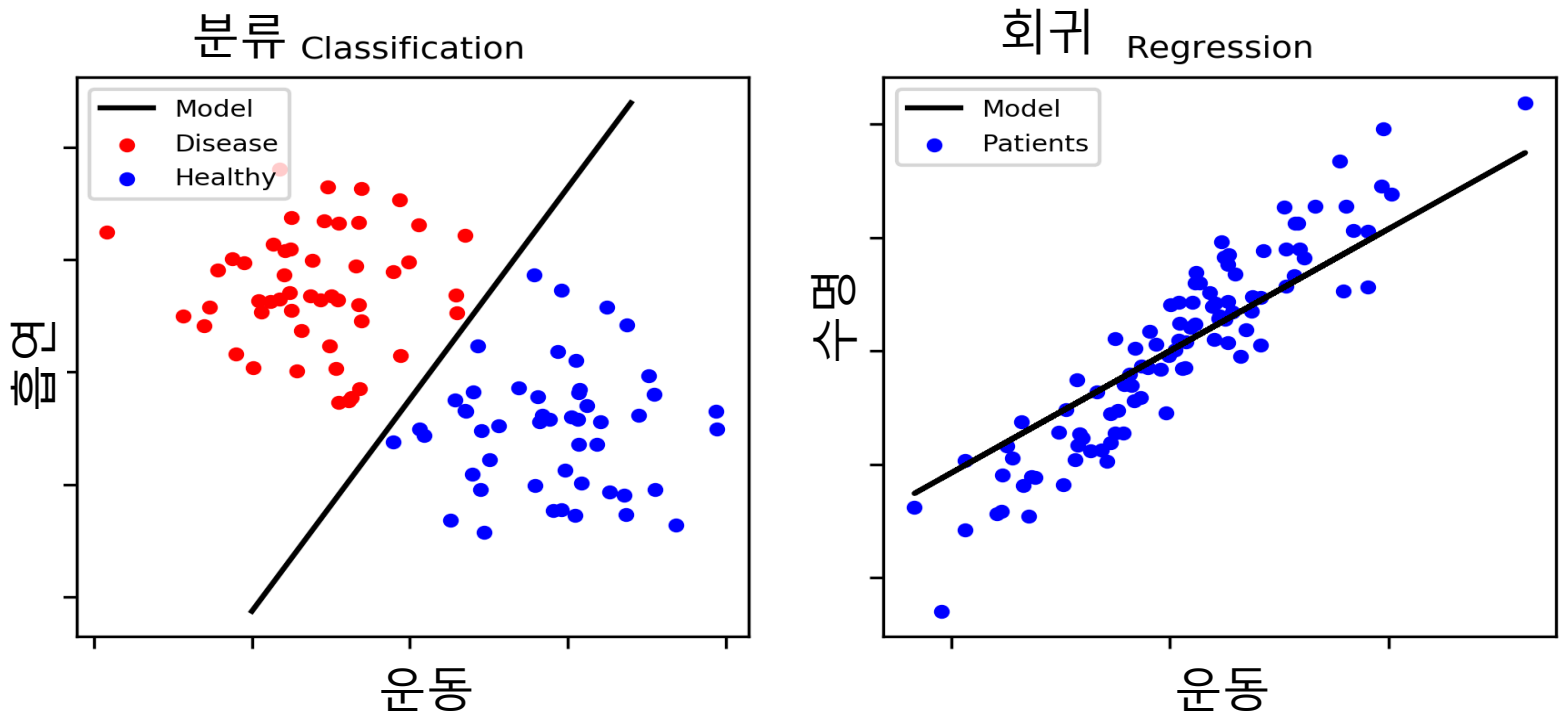
데이터 마이닝은 크게 **출력 변수**의 존재 여부에 따라 **지도학습과 자율학습으로** 나눌 수 있음

	지도학습 (supervised learning)	자율학습 (unsupervised learning)
의미	• 입력 데이터와 정답(Label)을 제공 받아 이를 통해 입력(독립)과 출력 (Label, 종속,타겟) 으로 매칭할 수 있는 규칙 생성 	• 외부에서 정답(Label)이 주어지지 않음 • 입력 데이터에서 패턴을 찾아내는 작업 
특징	• 출력 변수가 존재함	• 출력 변수가 존재하지 않음
분석 기법	• 의사결정나무, 회귀분석, 인공신경망, 판별분석 등	• 군집분석, 연관성 분석 등

지도학습 (supervised learning)

- 입력변수와 출력변수와의 관계를 학습하여 알려지지 않은 값을 예측하는 학습방법

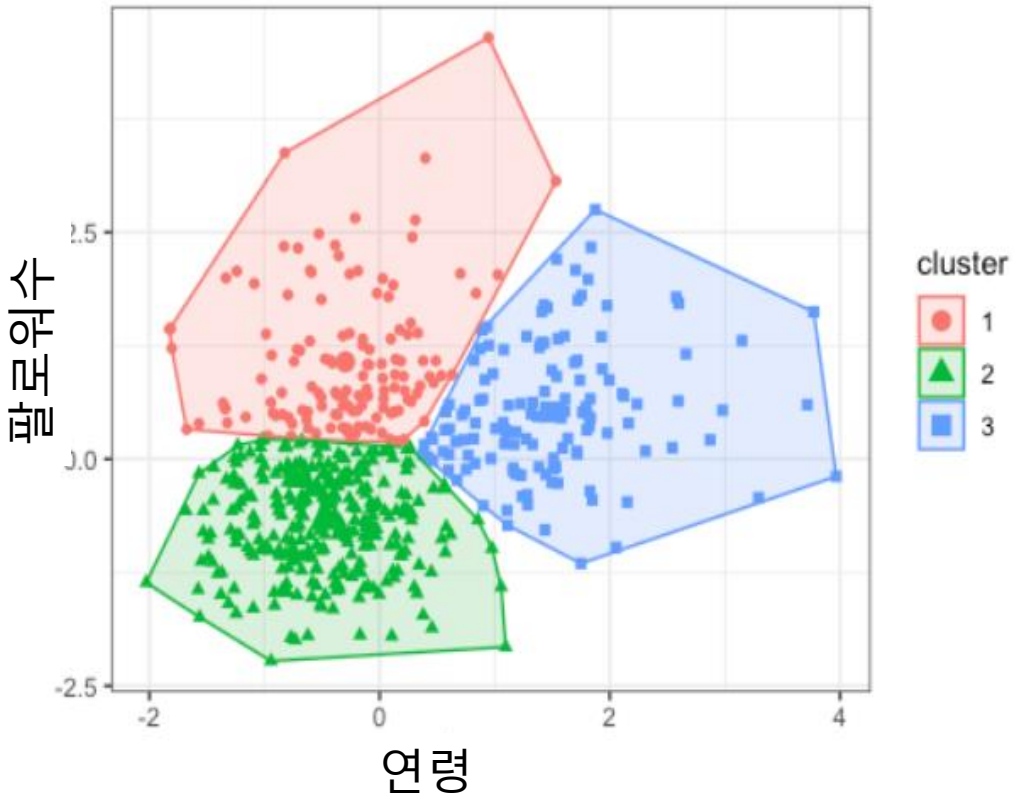
- 예. 분류 (2개 이상의 범주(카테고리, 클래스)), 회귀(실수 혹은 연속된 값) 등



자율학습 (unsupervised learning)

- 예측이 아닌, 정답이 없는, 데이터 간의 관계를 찾아내는 학습방법
- 예. 군집화 (유사성이 높은 데이터들의 그룹으로 나눔) , 연관성 분석(상품 구매와 같은 일련의 거래간의 규칙 발견) 등

군집화 Clustering



'언니의 파우치' 군집화 결과
생년월일, 팔로워수, 리뷰작성 수 등 11개 변수 이용

* 2016 중소기업 빅데이터 활용지원사업 우수사례집

2 데이터 마이닝 유형

분류 Classification

- 사전에 정의된 범주(클래스) 중 어디에 속하는지 결정
- 예. 고객 이탈 여부, 고객 등급 상/중/하, 품질 양호/보통/불량 등

회귀 Regression

- 주어진 입력-출력 쌍을 학습한 후 새로운 입력값이 들어왔을 때 출력값 예측
- 예. 고객 기여도, 고객 생애가치, 와인 품질 등

지도학습

군집 Clustering

- 데이터의 여러 속성들을 비교하여 유사한 특성을 갖는 몇 개의 그룹으로 나눔
- 예. 시장세분화 등

연관성 분석 Association Rule

- 한 패턴의 출현이 다른 패턴 출현을 암시하는 항목 간의 관계를 파악
- 예. 장바구니 분석 등

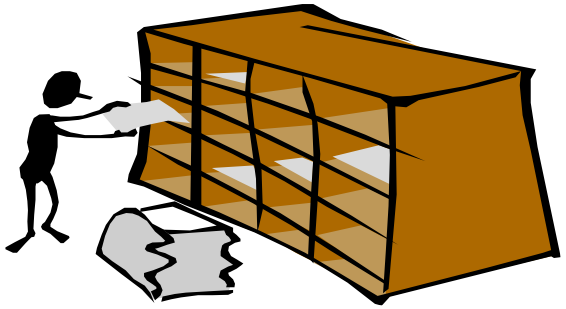
자율학습

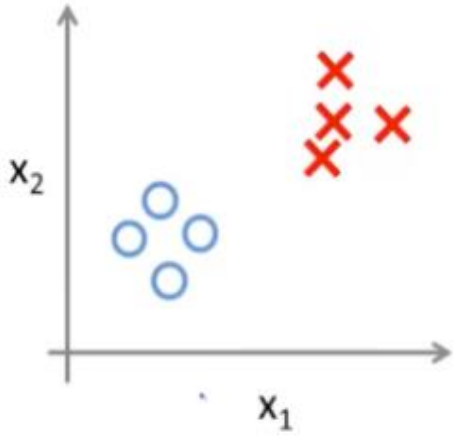
* 허명회, 이용구, "데이터 마이닝 모델링과 사례, 제2판," 한나래,

분류 (Classification)

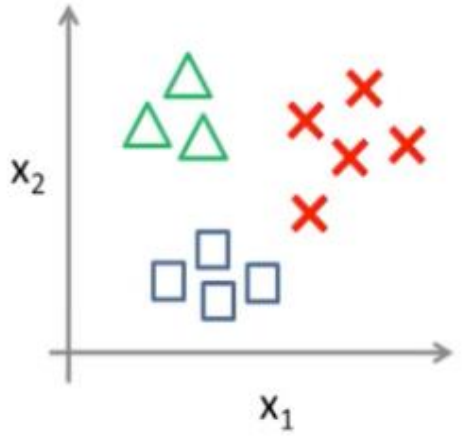
- 사전에 정의된 2개 이상의 범주 (클래스)로 분류할 수 있도록 해주는 모델(model)을 만드는 것
- 기존 데이터와 정답을 이용하여 모델을 만들고, 그 모델을 이용하여 새로운 데이터에 대해서 분류 수행

- 예
  - 고객 신용 등급 : 상 중 하 , 고객 충성도: 유지, 이탈
  - 클래스: 범주형 변수 (순서형 ordinal, 명목형 nominal)





이진 (binary) 분류



멀티클래스(multi-class) 분류

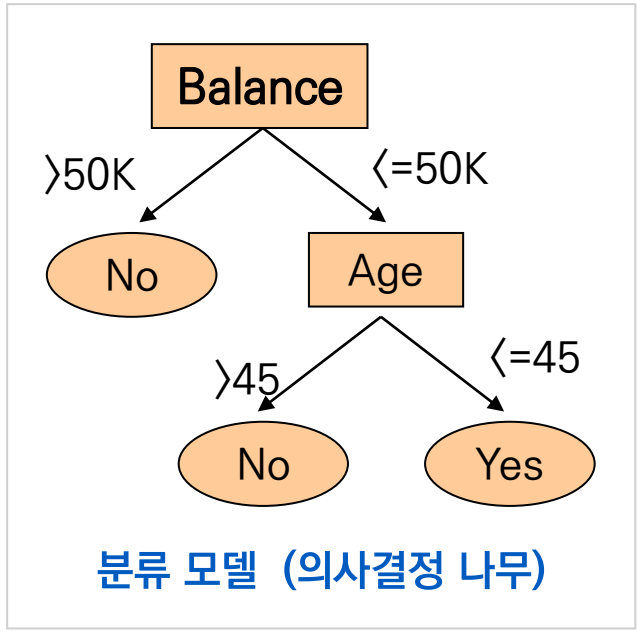
신용 위험 (credit risk) 관리 사례

- 의사결정나무 기법을 이용하여 고객의 채무불이행 여부(class: no, yes)를 예측
- {IF (condition) Then (class)} 규칙 도출
- 의사결정나무 모델은 설명 가능하여 이해하기 쉬움

Name 이름	Balance 잔고	Age 나이	Default 채무불이행
Mike	123,000	30	Yes
Mary	51,100	40	Yes
Bill	68,000	55	No
Jim	74,000	46	No
Mark	23,000	47	Yes
Anne	100,000	49	No

학습 데이터 (training data)

학습



분류 모델 (의사결정 나무)

Mark	88,000	40	?

테스트 데이터 (test data)

모델 적용

IF Balance > 50K, Then 'No'

회귀분석 (Regression)

- 입력 변수들을 이용하여 수치형인 출력 (타겟) 변수를 예측
- 예. 고객 기여도, 고객 생애가치, 소득 수준 등

기상조건에 따른 보르도(Bordeaux) 와인의 품질 예측

Y (와인 품질) = 12.145 + 0.00117 * X_1 (겨울 강우량)

+ 0.06140 * X_2(생장기 평균 기온)

average temperature during the growing season

− 0.00386 * X_3(추수기 강우량)

rainfall during the harvest months

Y: 보르도 와인 품질
(또는 평균 가격)

18개월 후

6개월 후

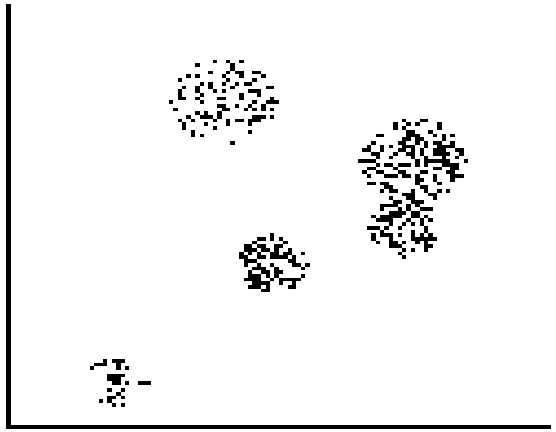
X: 1952~1980 보르도 날씨
(겨울/추수기 강우량,
생장기 기온 등)

Orley Ashenfelter
Professor of Economics at Princeton University
와인 평론가들이 시음해 보기도 전에 미래 가치 예측 가능

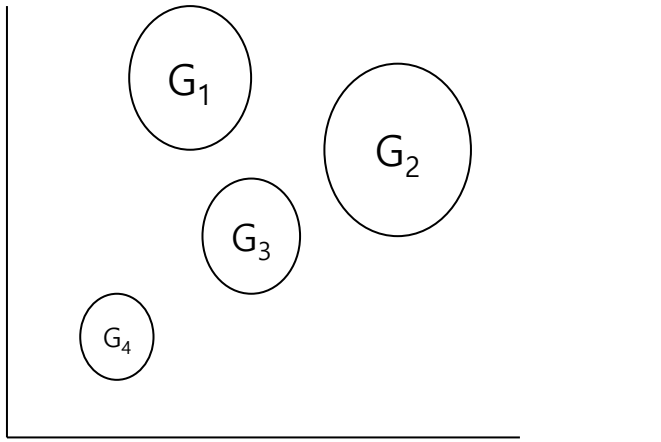
군집화 (Clustering)

- 이질적인 모집단으로부터 다수의 동질적인 하위 군집으로 세분화하는 작업
- 분류 기준 없이 비슷한 것끼리 묶은 후, 군집 결과에 의미를 부여하는 것은 사용자의 몫임
- 다른 데이터 마이닝 기법을 적용하기 전 사전 작업으로 활용

- 예. 유사한 구매행동을 보이는 고객들이 한 그룹이 되도록 시장 세분화



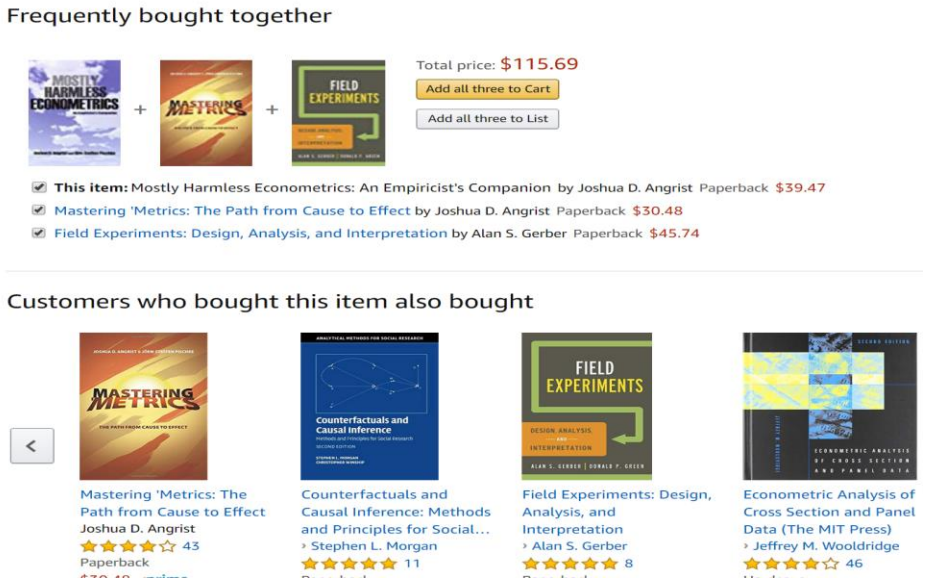
이질적인 모집단



4개의 동질적인 하위그룹(cluster)

연관성 분석 (Association Rules)

- 동시에 발생하는 데이터로부터 변수간의 규칙 생성
- 대표적인 사례는 장바구니 분석으로 동시 구매 가능성이 높은 품목에 대한 규칙을 찾는 방법 (기저귀와 맥주)



아마존: 함께 판매된 제품간의 관계를 기반으로 관련 제품 추천

규칙번호	제품 1	제품 2	지지도	신뢰도	향상도
1	간호사 스프링 집게 가위 집게줄	간호사 주머니 널스포켓 8 color	0.091	0.374	2.333
2	간호사 스프링 집게 가위 집게줄	널쓰리 포켓 5 컬러 (대용량널스포켓)	0.053	0.218	2.083
3	널쓰리 포켓 5 컬러 (대용량널스포켓)	간호사 스프링 집게 가위 집게줄	0.053	0.508	2.083
4	의료용 가위 일자형 간호사 가위	간호사 스프링 집게 가위 집게줄	0.052	0.563	2.309
6	똑딱이 네임펜 3color/노크식 네임펜	간호사 스프링 집게 가위 집게줄	0.050	0.523	2.144
7	간호사 스프링 집게 가위 집게줄	똑딱이 네임펜 3color/노크식 네임펜	0.050	0.206	2.144
8	간호사 테이프걸이 릴홀더 캐릭터 / 아크릴	간호사 스프링 집게 가위 집게줄	0.040	0.453	1.859
9	간호사 테이프걸이 릴홀더 캐릭터 / 파스텔	간호사 스프링 집게 가위 집게줄	0.040	0.525	2.151
10	간호사 테이프걸이 릴홀더 캐릭터 / 일반	간호사 스프링 집게 가위 집게줄	0.036	0.602	2.469
		⋮			
112	간호사 스프링 집게 가위 집게줄	간호화 (뽀너스 국내생산 가볍고 푹신한 시그니처 운동화형 간호화 남여공용)	0.010	0.042	0.499

[제품간 연관성 분석 결과]

⇒ 뽀너스는 고년차 간호사들의 업무경험에 따라 연관구매가 빈번한 것으로 판단되는 '간호아이디어 제품'을 중심으로 세트제품을 기획하고, 단체구매 조직문화로 연관구매가 잘 일어나지 않는 '간호화'와 추가구매가 가능한 제품을 발굴하기로 했다. 또한 자사몰 2차 리뉴얼시 연관구매가 빈번히 발생하는 제품들 간의 이동거리를 최소화하는 방향으로 자사몰 기능을 개선하는 방안을 고려 중이다.

뽀너스: 간호사를 위한 온라인 쇼핑몰

7 데이터 마이닝 유형: 기타

이상치 또는 특이값 (anomaly) 탐지

- 다른 데이터 포인트들과 크게 차이가 나는 데이터 식별
- 신용카드 사기탐지(fraud detection)

텍스트 마이닝

- 입력 데이터가 문서, 메시지, 이메일, 웹페이지 형식과 같은 텍스트인 경우

시계열 예측

- 동일한 변수의 과거 값들을 기준으로 하여 미래 값을 예측
- 평균화(averaging) 또는 평활화(smoothing)

특징 선택 (feature selection)

- 데이터의 속성(feature)들을 매우 중요한 몇 개의 속성들로 줄이는 과정

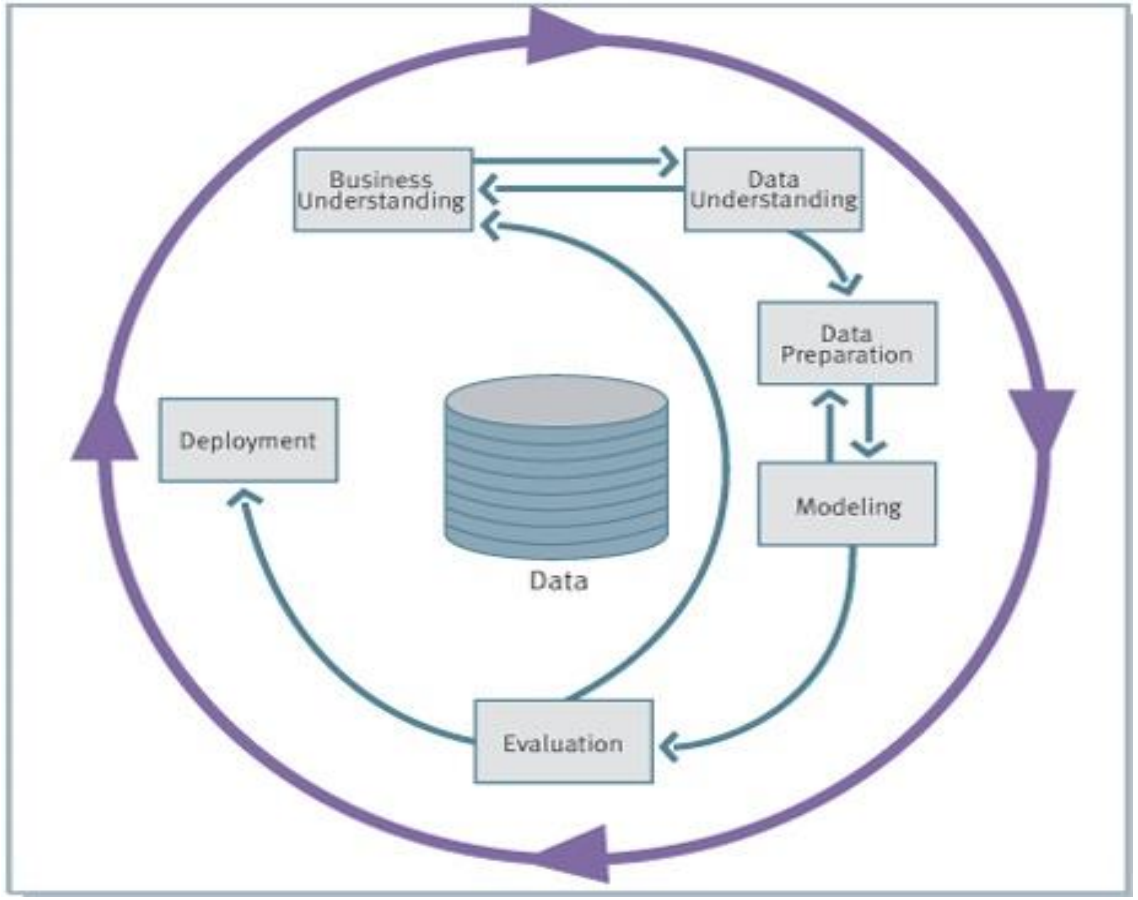
Ⅳ. 데이터 마이닝 프로세스

데이터 마이닝 프로세스의 일반적인 순서 (대표적인 데이터 마이닝 프레임워크 : CRISP-DM)

1. 비즈니스 이해 (Business Understanding)
2. 데이터 이해 (Data Understanding)
3. 데이터 준비 (Data Preparation)
4. 모델링 (Modeling)
5. 모델 평가 (Evaluation)
6. 전개 (Deployment, 실제 시스템에 사용 배포)

- 데이터 마이닝 프로세스는 일회적이 아닌 반복적으로 수행되어야 함
- 비즈니스와 데이터에 대한 이해, 데이터 준비에 시간을 더 많이 투자해야 함

CRISP-DM(Cross Industry Standard Process for Data Mining)



전체 프로세스 중 80% 정도 비중을 차지 함

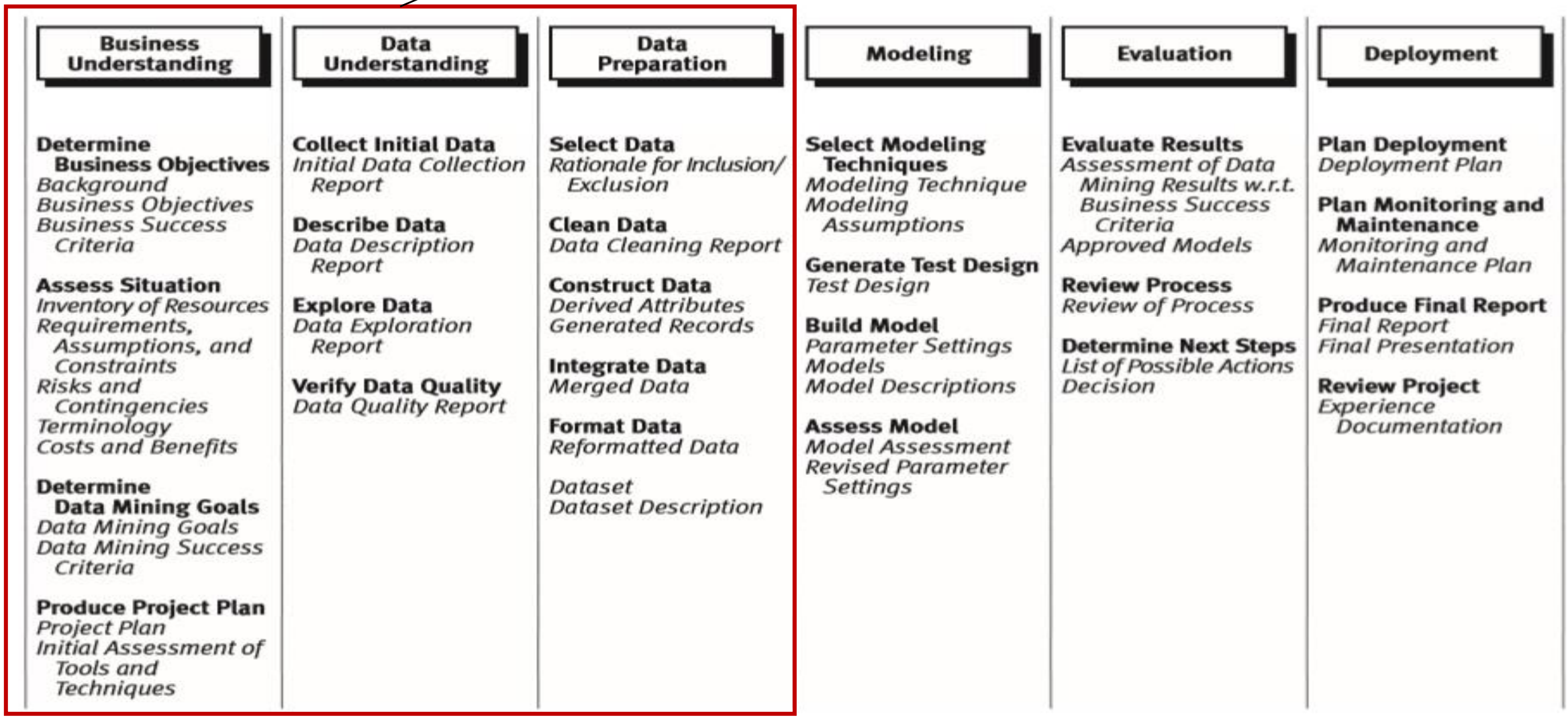
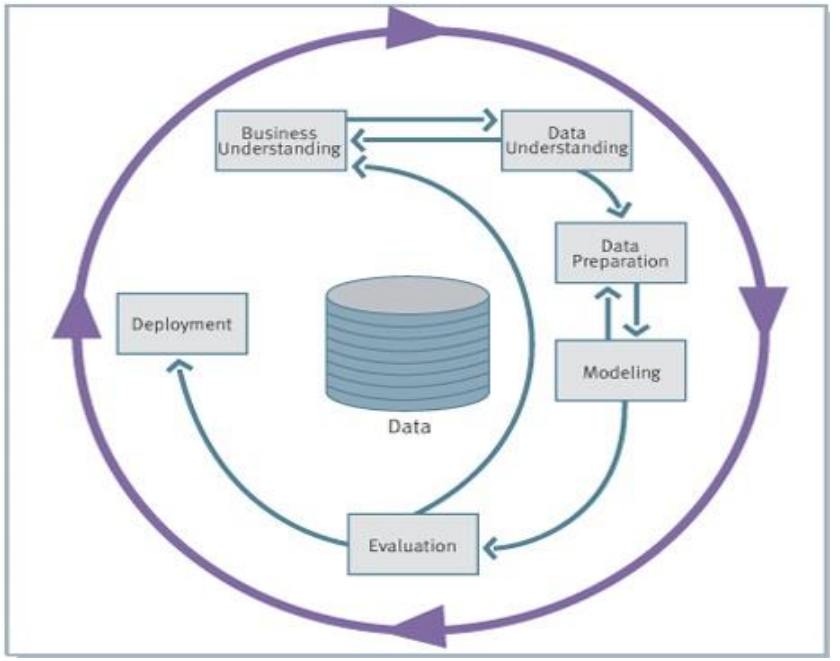


Figure 3: Generic tasks (bold) and outputs (italic) of the CRISP-DM reference model

2 데이터 마이닝 프로세스 단계



1. 현장 업무 이해 (Business Understanding)

- 비즈니스 문제를 데이터 과학의 문제로 전환하는 단계
- 정확히 비즈니스 문제를 이해하여 데이터 과학으로 어떻게 해결하는지를 정의하는 단계 (반드시 현업 책임자와 커뮤니케이션 필요)

2. 데이터 이해 (Data Understanding)

- 비즈니스 문제를 해결하기 위한 데이터 획득여부와 가용여부 판단하는 단계
- 분석에 적절한 데이터 구조를 파악하는 단계

3. 데이터 준비 (Data Preparation)

- 데이터 정제, 새로운 데이터 생성, 데이터 업데이트 등 원천 데이터를 분석에 적합한 형태로 변환하는 단계로 가장 시간을 많이 투자해야 하는 단계
- 예. 결측치 (missing value), 데이터 유형 및 전환 (data type and conversion), 특이값 (outlier) 처리 등
- 텍스트,이미지 등의 사용 증가로 인해 비정형데이터를 다룰 수 있는 관련 알고리즘 개발이 활발해 짐

4. 모델링 (Modeling)

- 비즈니스 문제를 해결하기 위해 가장 적합한 기법 선택 및 유의미한 모형 구축 단계
- 예. 의사결정나무, 인공신경망, 군집 분석 등

5. 평가 (Evaluation)

- 분석 모델의 결과를 평가하고, 다음 단계로 넘어가도 되는지 모델의 신뢰성을 확인하는 단계임
- 모델의 정확성 뿐 아니라 비즈니스에서의 적합성도 살펴봐야 함
- 모델 생성 뿐 아니라 시스템화 후의 모델 평가도 포함 (모델의 주기 파악)

6. 전개 (Deployment)

- 검토가 끝난 모델을 실제 현업에 적용하는 단계
- 완성된 마이닝 프로세스를 자동화/시스템화 하는 과정
- 모델링 까지는 데이터 과학팀, 시스템화 부터 운영 및 유지 등은 개발팀의 책임

다른 데이터 마이닝 프레임워크
- SEMMA(Sample, Explore, Modify, Model, Assess)
- DMAIC(Define, Measure, Analyze, Improve, Control)
- KDD(Knowledge Discovery in Databases, Selection, Preprocessing, Transformation, Data Mining, Interpretation, and Evaluation framework)

3 정리

데이터 마이닝 정의

- 데이터 속의 **유용한 패턴(규칙, 관계)**을 찾고 이를 일반화하는 프로세스
- 관측 데이터에 적합한 모델을 구축하는 과정

데이터 마이닝 유형

- 지도학습과 자율학습의 차이
- 분류, 회귀, 군집화, 연관성 분석, 텍스트 마이닝 등

데이터 마이닝 프로세스

- 최적의 모델과 결과를 얻기 위해서는 비즈니스 이해, 데이터 이해, 데이터 준비, 모델링, 평가, 전개의 단계를 반복적으로 수행