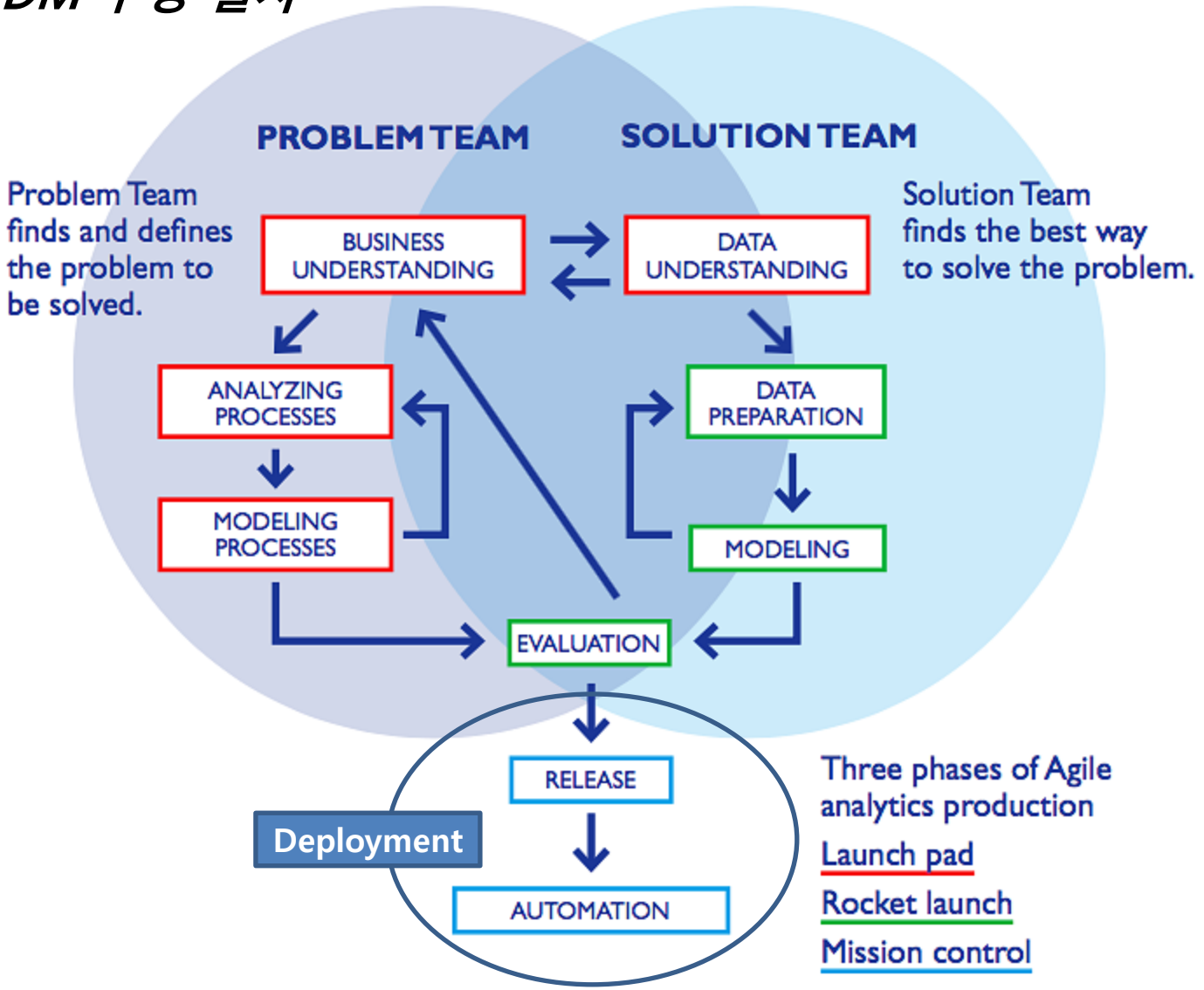
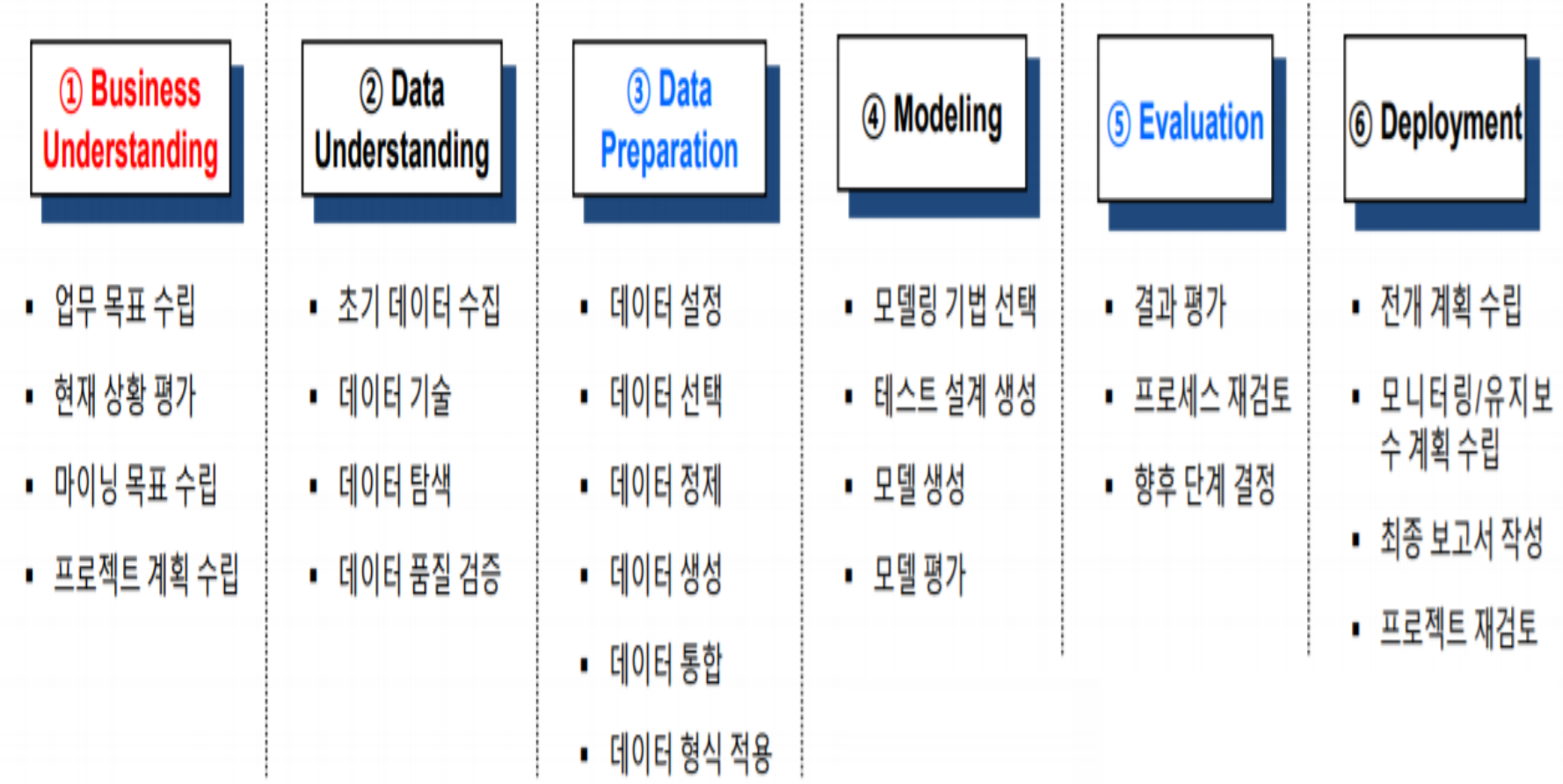
빅데이터
프로젝트 방법론


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

CRISP-DM 수행 절차



CRISP-DM의 프로세스 단계



마이닝 단계	General Task	Output	설명
Business Understand	Determine Business Obejectives	Success Criteria	비즈니스 관점 이해
	Assess Situation	Costs and Benefits	영향평가 및 상황조사
	Determine Data Mining Goals	Goals and Criteria	기술관점 목표수립
	Produce Project Plan	Initial Assessment	프로젝트 세부계획 수립
Data Understand	Collect Initial Data	Initial Data Report	초기데이터 수집/확보
	Describe Data	Describe Report	수집데이터 특성 확인
	Explore Data	Explore Report	목표 부합 데이터 추출
	Verify Data Quality	Quality Report	데이터 품질 검사/검증

마이닝 단계	General Task	Output	설명
Data Preparation	Select Data	Rationale	분석대상 데이터선별
	Clean Data	Cleaning Report	데이터정제, 품질확보
	Construct Data	Attributes/Records	분석데이터 구조화
	Integrate Data	Merged Data	데이터통합
	Format Data	Reformatted Data	모델적합목적 형식보완
Modeling	Select Modeling Techniques	Tech and Assumption	데이터모델링 기법선택
	Generate Test Design	Test Design	품질검증/유효성검사
	Build Model	Parameter Setting	유효 모델링 생성

마이닝 단계	General Task	Output	설명
Evaluation	Evaluate Results	Approved Models	측정 / 모델링 결과 평가
	Review Process	Review Report	중요항목 도출/확인
	Determine Next Steps	List of Actions	다음단계 결정
Deployment	Plan Deployment	Plan Report	전개전략 수립
	Plan Monitoring and Maintenance	Detailed Plan	지속/유지전략 수립
	Produce Final Report	Final Report	최종보고서 작성
	Review Project	Final Documentation	최종보고서 검토

3 단계 – 데이터 준비

- **데이터 선택(Select your data)**
  - 이것은 분석을 위해 사용할 데이터를 결정하는 프로젝트의 단계
  - 이러한 결정을 내리는 데 사용할 수 있는 기준에는 데이터 마이닝 목표와의 관련성, 데이터 품질 및 데이터 볼륨 또는 데이터 유형에 대한 제한과 같은 기술적 제약이 포함
  - 데이터 선택은 테이블의 레코드(행) 선택뿐만 아니라 속성(열)의 선택을 다룸
  - 포함/배제의 이론적 근거(Rationale for inclusion/exclusion)
    - ✓ 데이터를 포함/제외하고 이러한 결정에 대한 이유를 설명합니다.

- **데이터 정제(Clean your data)**
  - 이 작업에는 선택한 분석 기술에 필요한 수준으로 데이터 품질을 높이는 작업이 포함
  - 여기에는 데이터의 깨끗한 부분 집합 선택, 적절한 기본값의 삽입 또는 모델링을 통해 누락된 데이터를 추정하는 등보다 공격적인 기법이 포함될 수 있음
  - 데이터 정제 보고서(Data cleaning report)
    - ✓ 데이터 품질 문제를 해결하기 위해 취한 조치와 조치를 설명
  - 정제 목적으로 만들어진 데이터의 변형과 분석 결과에 미칠 수 있는 영향을 파악

3 단계 – 데이터 준비

- **필수 데이터 구성(Construct required data)**
  - 이 작업에는 파생 된 특성 또는 전체 새 레코드의 생성 또는 기존 특성의 변환 된 값과 같은 구성적인 데이터 준비 작업이 포함
  - 파생 속성(Derived attributes) – 동일한 레코드의 하나 이상의 기존 속성으로 구성된 새로운 속성입니다. 예를 들어 length 및 width 변수를 사용하여 영역의 새 변수를 계산할 수 있음
  - 생성 된 레코드(Generated recordㄴs) – 완전히 새로운 레코드의 생성을 설명
  - 예를 들어 작년에 구매하지 않은 고객에 대한 레코드를 작성해야 할 수 있습니다.
  - 원시 데이터에 그러한 레코드를 가질 이유는 없었지만, 모델링 목적을 위해 특정 고객이 구매를하지 않았다는 사실을 명시 적으로 나타내는 것이 합리적 일 수 있음

- **데이터 통합(Integrate data)**
  - 이는 여러 데이터베이스, 테이블 또는 레코드에서 정보를 결합하여 새로운 레코드 값을 만듬.
  - 병합된 데이터(Merged data)
    - ✓ 병합 테이블은 동일한 오브젝트에 대한 서로 다른 정보가 있는 두개 이상의 테이블을 결합하는 것을 말합니다.
    - ✓ 예를 들어, 소매 체인에는 각 매장의 일반적인 특성(예 : 매장 공간, 몰 유형)에 대한 정보가 있는 테이블 하나, 요약된 판매 데이터가 있는 또 다른 테이블(예 : 이익, 전년 대비 매출 변동율) 및 주변 지역의 인구 통계에 관한 정보 테이블을 합칩니다.
    - ✓ 각 테이블에는 각 상점에 대해 하나의 레코드가 들어 있습니다.
    - ✓ 이 테이블은 소스 테이블의 필드를 결합하여 각 상점에 대해 하나의 레코드가 있는 새 테이블로 병합될 수 있습니다.
  - 집계(Aggregation)
    - ✓ 집계는 여러 개의 레코드나 테이블의 정보를 요약하여 새 값을 계산하는 작업을 말합니다.
    - ✓ 예를 들어 각 구매에 대해 하나의 레코드가 있는 고객 구매 테이블을 구매 수, 평균 구매 금액, 신용 카드로 부과된 주문 비율 등의 필드가 있는 각 고객에 대한 레코드가 있는 새 테이블로 변환합니다.