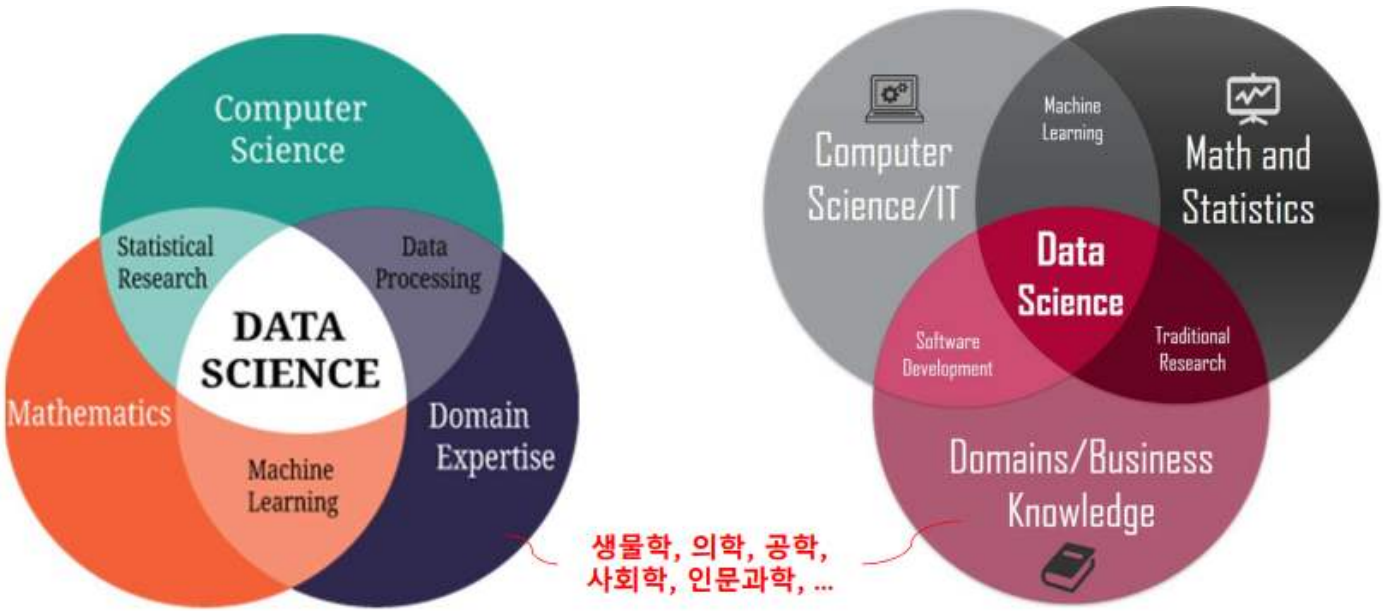
## Data Science

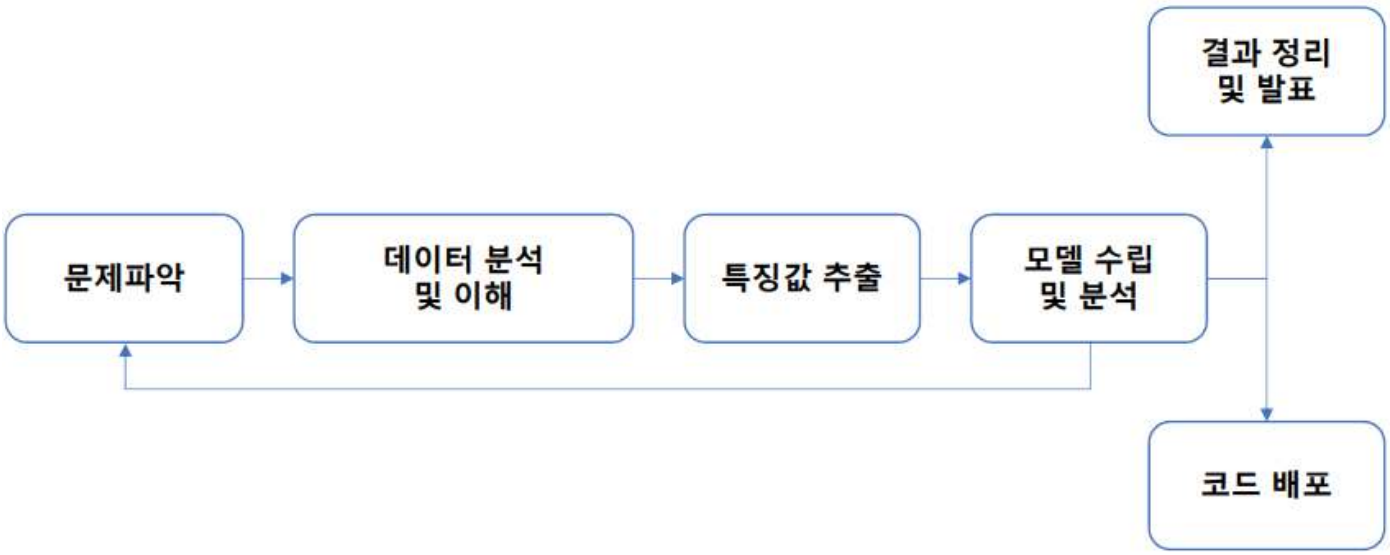
- 데이터를 수집, 처리, 분석하여 유의미한 통찰(insignt)을 이끌어내는 것
- 제조업, 금융, 전자 상거래, 의료 등 다양한 분야에서 사용되고 있음

- 문제 정의 및 해결 방법 탐색 능력
- 적절한 데이터 수집 및 처리
- 과학적 분석 처리 방법, 프로세스, 알고리즘, 모델, 시스템 등 지식 / 개발 역량
- 분석 결과 이해 능력
- 설득력 있는 커뮤니케이션 스킬




- Computer Science: 대용량 데이터 처리, 데이터 전처리, 분석 모델 개발, ...
- Mathematics: 데이터의 통계적 가치 이해, 분석 모델 알고리즘, ...
- Domain Knowledge: 데이터의 인문학적 가치 이해, ...

## Data Science Roadmap



### 문제 파악

- 문제 정의
- 프로젝트 요구 사항 정의 및 목표 선정

### 데이터 분석 및 이해

- 문제 해결을 위한 데이터 수집
- 수집한 데이터 이해: 데이터 크기, 자료형, 이상치, 통계적 특징 등
- 데이터 전처리: 중복 등 불필요한 데이터 제거, 결측치 처리
- 데이터 분석: 데이터 통계 분석, 시각화 등을 이용해 데이터 특성/의미 분석

### 특징값 추출

- 데이터 특성을 고려하여, 모델 이용에 적합한 형태로 값 변형/추출

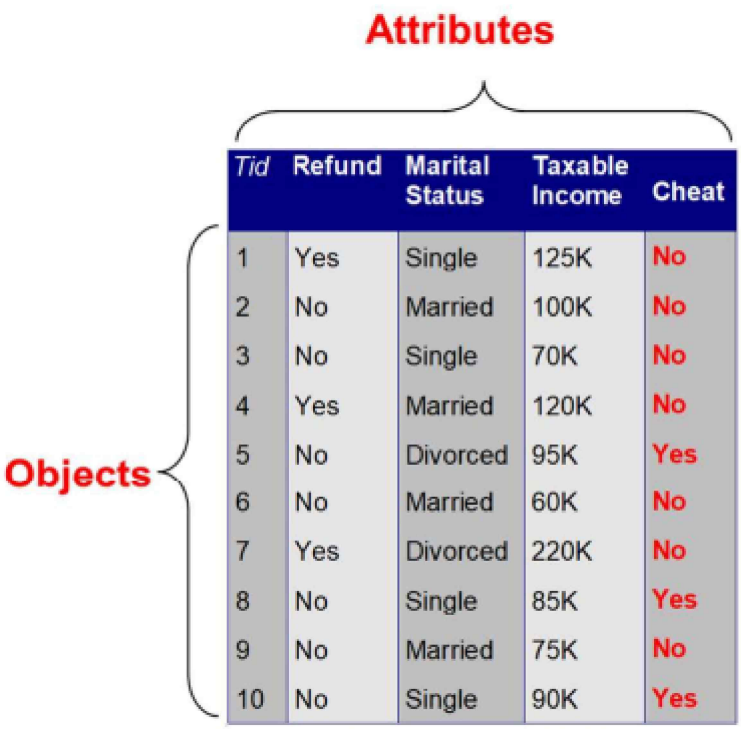
### 모델 수립 및 분석

- 문제 해결에 적합한 모델 적용 및 결과 분석
- 분류(Classification): 데이터 샘플이 어떤 카테고리에 속하는지 추론
- 군집화(Clustering): 구체적인 특성을 공유하는 집단(군집)들로 구분
- 연관성(Association): 여러 사건들 간의 관계 정의
- 연속성(Sequencing): 가까운/특정 기간 동안 일어나는 사건들 간의 관계 정의
- 예측(Forecasting): 데이터의 패턴을 이용하여 미래 값 예측

## Data

데이터를 모델링하기 위해선 컴퓨터가 인식할 수 있는 형태로 변형하는 과정이 필요함

- 정형(structed) 데이터: 형식과 구조가 명확히 정의된 필드에 기록된 데이터. csv, database 등
- 반정형(semi-structed) 데이터: 형식과 구조는 있으나 변경될 수 있는 데이터. html, xml, json 등
- 비정형(unstructed) 데이터: 값 추출 규칙이 명확하지 않은 데이터. 텍스트, 음성, 영상 등




| Tid | Refund | Marital Status | Taxable Income | Cheat |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 95K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 75K | No |
| 10 | No | Single | 90K | Yes |

- 데이터 개체(object): case, sample, observation, instance, record, pattern, vector 등
- 데이터 속성(attribute): variable, feature 등

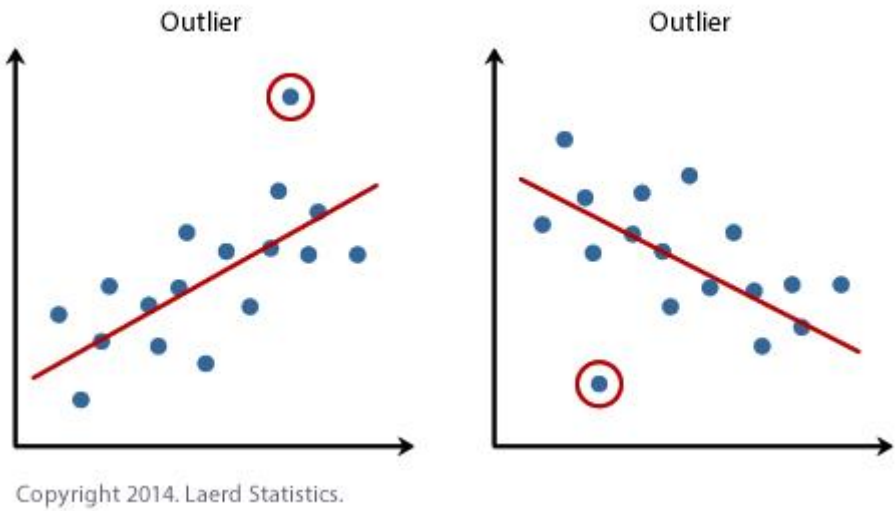
**데이터 속성**

- 양적 자료(Quantitative, Numeric): 수치로 관측된 값
- 질적 자료(Qualitative, Categorical): 범주(카테고리)나 순서 형태의 값

**데이터 품질 이슈**

- 실제 환경에서 완벽한 데이터를 수집하는 것은 불가능함
- 잡음(noise): 측정 과정에서 무작위로 에러가 발생하는 것
- 바이어스(bias): 측정 장비에 포함된 시스템적인 변동. 데이터 전체 품질에 영향을 미침. 영점 조절되지 않은 체중계 등
- 중복(duplicate): 동일한 개체가 데이터 내에 여러번 나타나는 경우. 불일치가 있을 시 추가 작업 필요
- 이상치(outlier): 특히 예외적으로 다른 성질을 보이는 일부 데이터
- 결측치(missing value): 누락된 데이터

데이터 품질 이슈를 다소 해소하고 모델 성능 향상을 위해 전처리(preprocessing) 과정을 거침




| | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| 0 | NaN | 41.0 | 6.984127 | 1.023810 | 322.0 |
| 1 | 8.3014 | 21.0 | 6.238137 | 0.971880 | 2401.0 |
| 2 | NaN | 52.0 | 8.288136 | 1.073446 | 496.0 |
| 3 | 5.6431 | NaN | 5.817352 | NaN | 558.0 |
| 4 | NaN | 52.0 | 6.281853 | 1.081081 | 565.0 |
| 5 | 4.0368 | NaN | 4.761658 | 1.103627 | 413.0 |
| 6 | 3.6591 | 52.0 | 4.931907 | 0.951362 | 1094.0 |
| 7 | NaN | 52.0 | 4.797527 | 1.061824 | 1157.0 |
| 8 | NaN | 42.0 | 4.294118 | 1.117647 | 1206.0 |
| 9 | 3.6912 | 52.0 | 4.970588 | 0.990196 | 1551.0 |

- 결측치가 있는 object나 attribute 삭제
- 결측치를 대신하여 0 등 임의의 값 사용
- 결측치가 있는 attribute의 값을 이용하여 결측치 값을 추정하여 사용(평균, 중앙값, 최빈값 등)
- ...