/* elice */

문과생을 위한 머신러닝

3주차: 머신러닝을 위한 데이터 이해하기

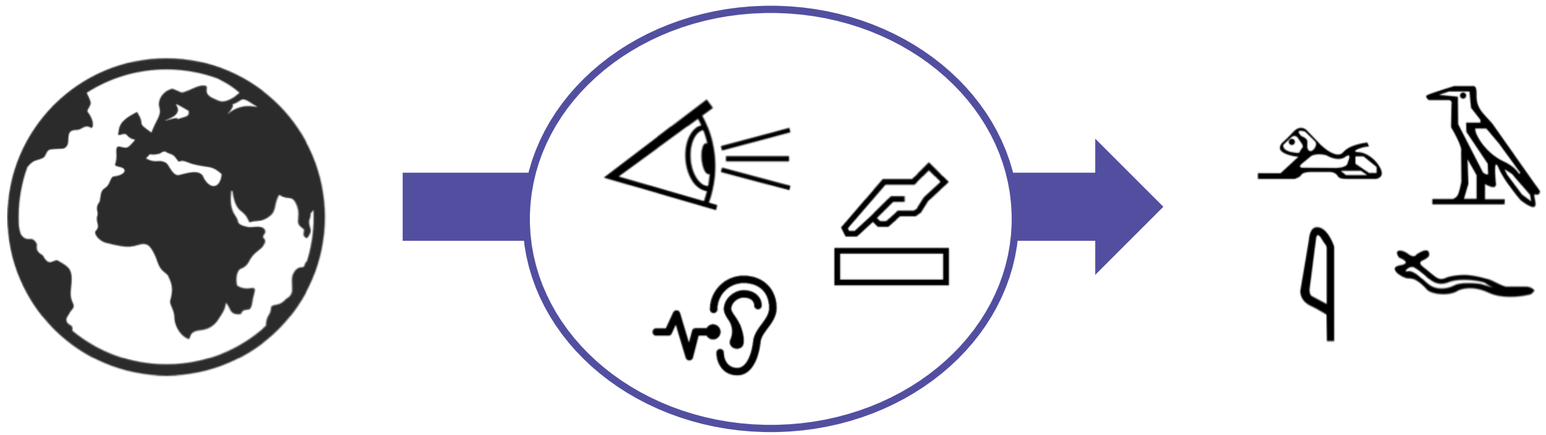
David Oh 선생님

목차

1. 머신러닝을 위한 핵심개념 살펴보기

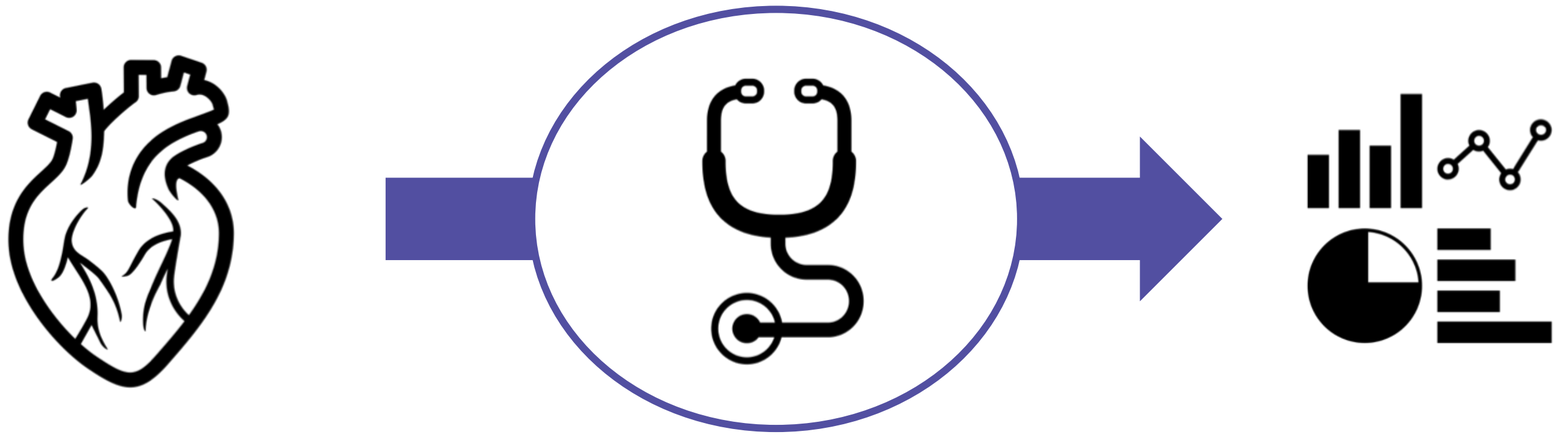
Data

현실 세계의 어떤 현상을 관찰하여 기록한 것

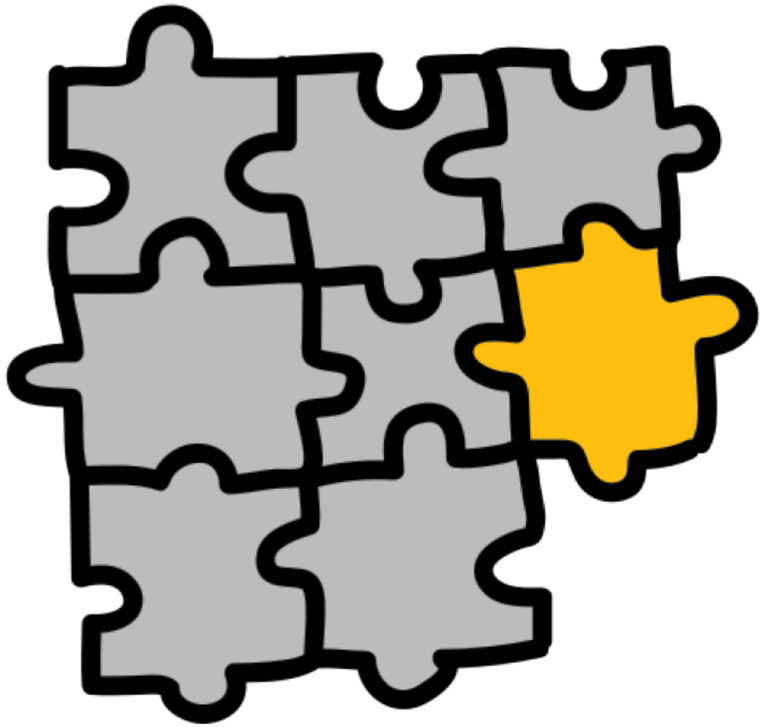


Heart Data

심장 상태를 관찰하여 기록한 것



Data



데이터의 특징

- Facts
- No meaning
- Representation of real world

Feature

Feature는 데이터(data)를 **컴퓨터가 이해할 수 있도록**

수치(numeric) 또는 디지털(digitized)로

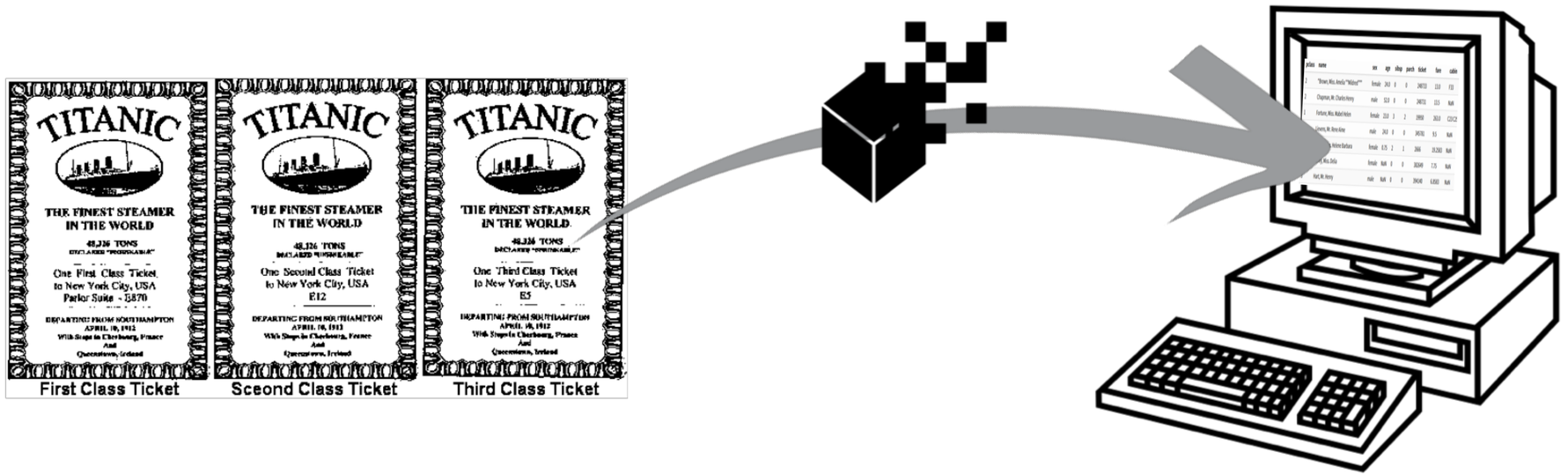
표현/**표상**(representation)한 것

Titanic Feature

타이타닉호 티켓을

디지털 형태로
컴퓨터가 이해하도록

표현/표상한다



Synonym of “Feature”

- Independent Variable
- Explanatory Variable
- Predictor
- Input
- Attribute

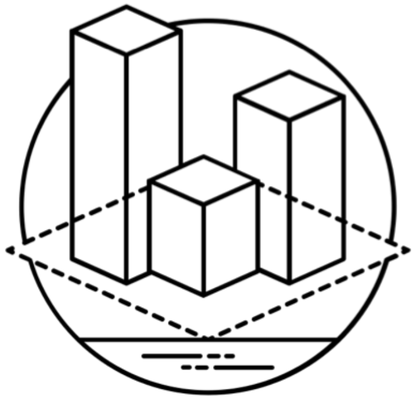
Target

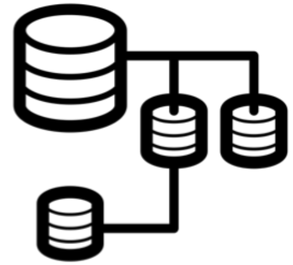
예측하려는 목표

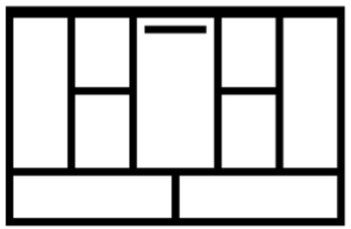
Model ?

실제의 무엇을 더 작게 추상화된 형태로 표현한 것

모형 또는 본보기



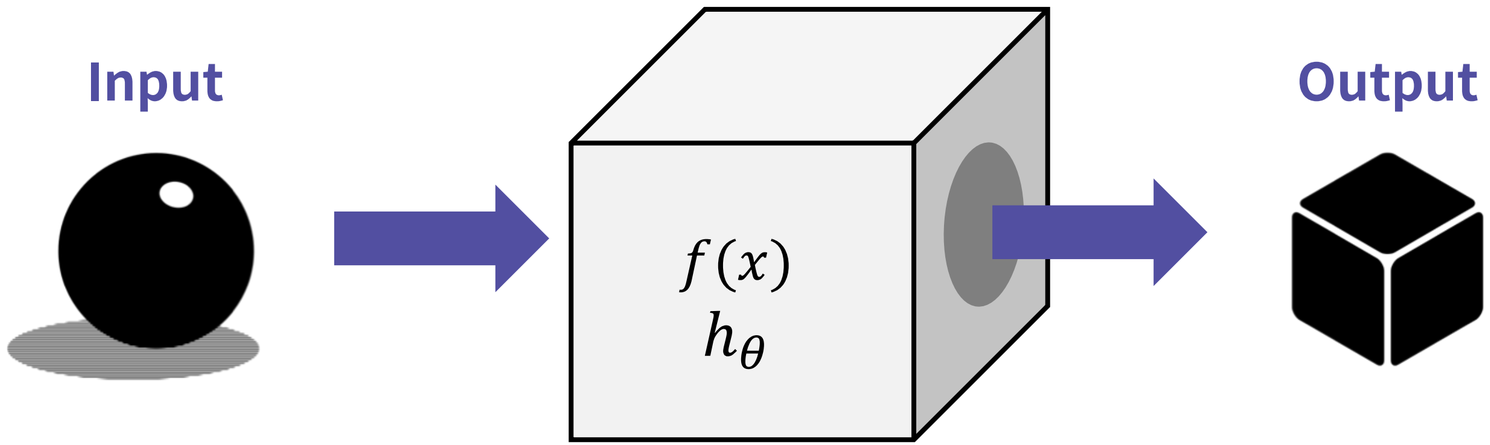




Machine Learning Model

어떠한 문제를 해결하기 위해 수립한 **가설**을

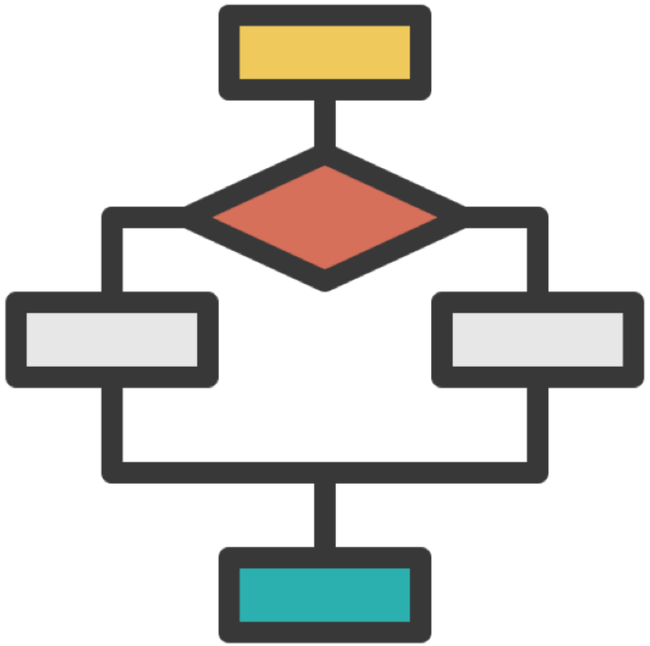
논리적, 수학적 **함수식**의 형태로 표현한 것



Algorithm

입력된 자료를 바탕으로 원하는 **결과**를 유도하기 위해

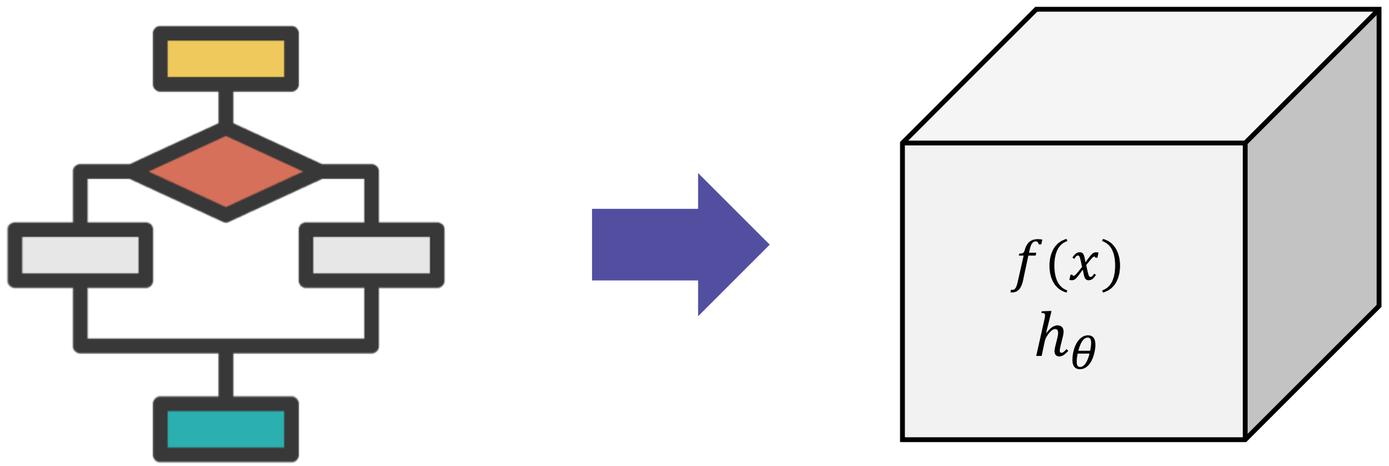
일련의 논리적인 순서와 절차를 **규칙**화한 것



Machine Learning Algorithm

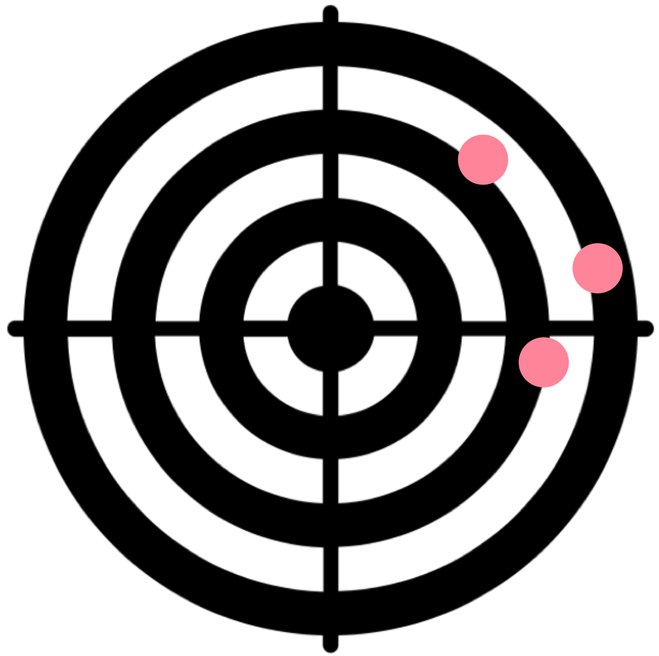
Model이 어떠한 문제를 해결하기 위한 **함수식**이라면

Algorithm은 그 함수식을 만들어내는 **일련의 절차, 규칙**



Loss, Cost, Error

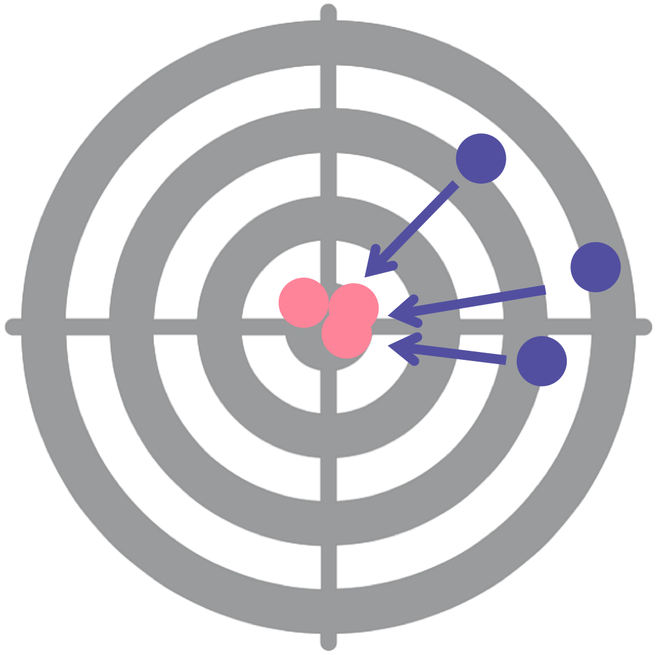
예측 목표로부터 예측 결과의 **오차**



Learning

예측 목표로부터 예측 결과의 **오차를 최소화**하는

함수식을 찾아내는 과정



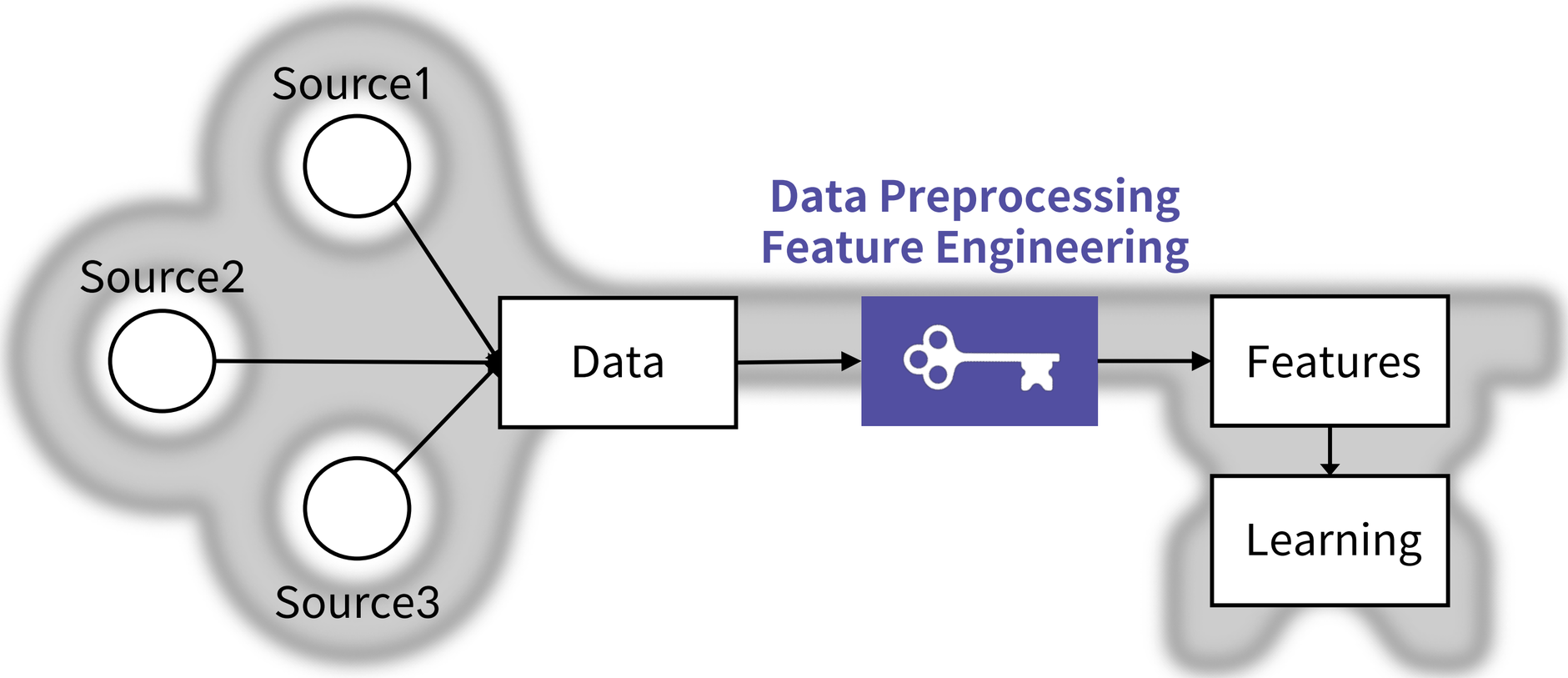
2. 머신러닝을 위한 데이터 준비

Think First,

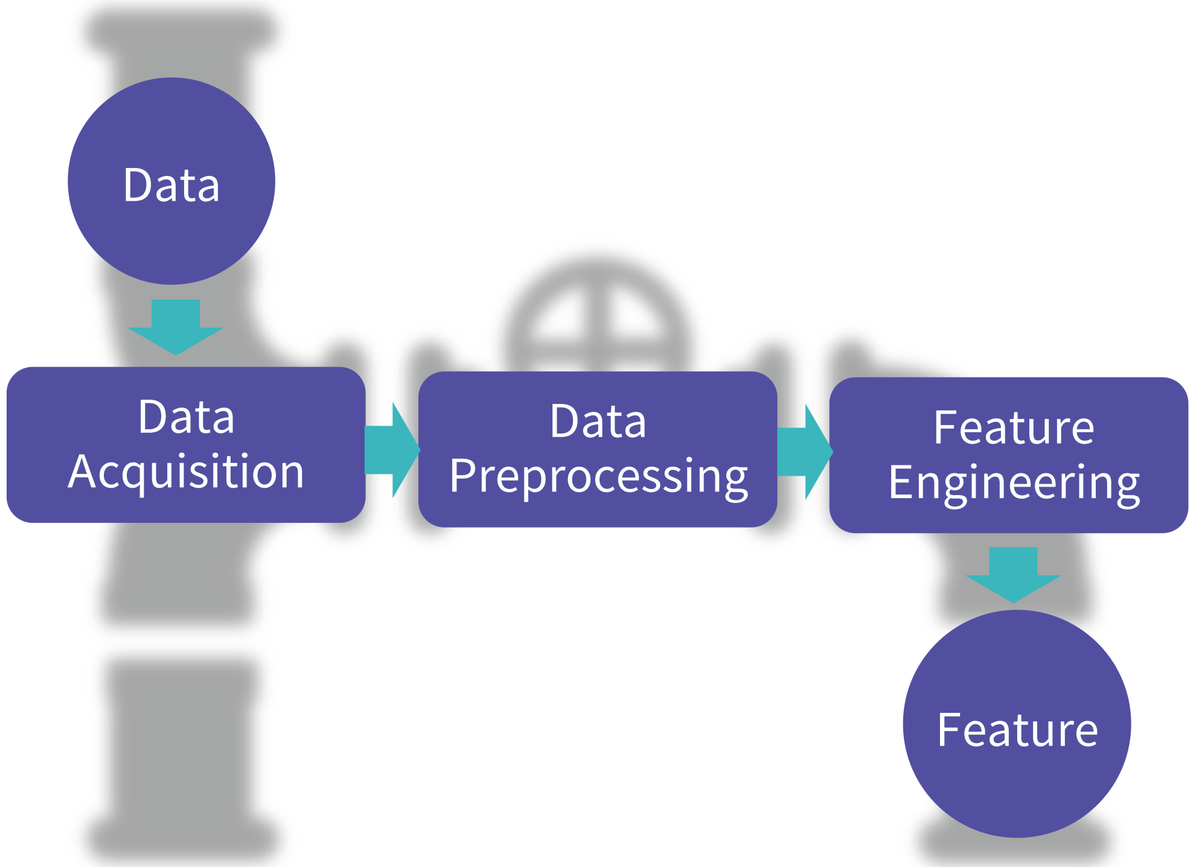
Garbage Input

Garbage Output

Data Preparation



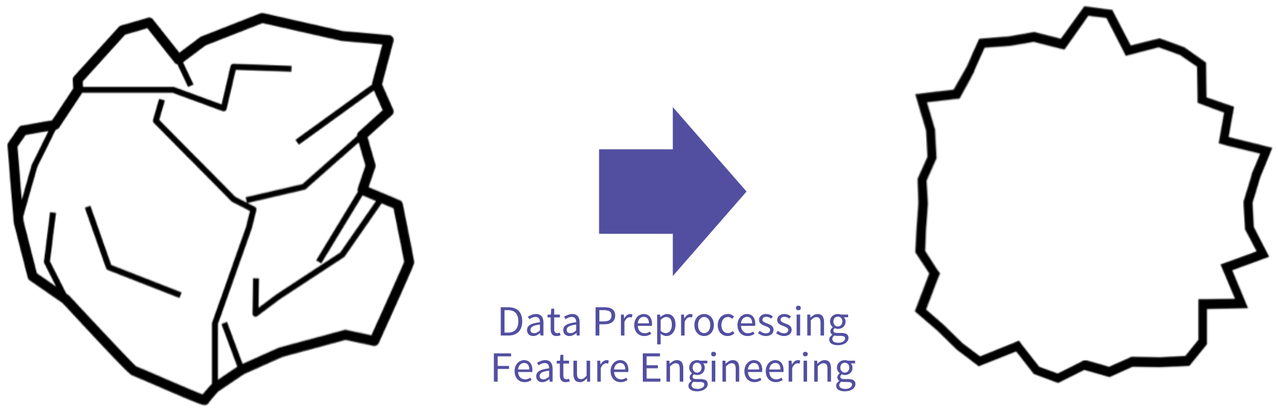
Data Preparation Pipeline



머신러닝 프로젝트 성공의 열쇠

대다수의 Data Preprocessing과 Feature Engineering 기법은

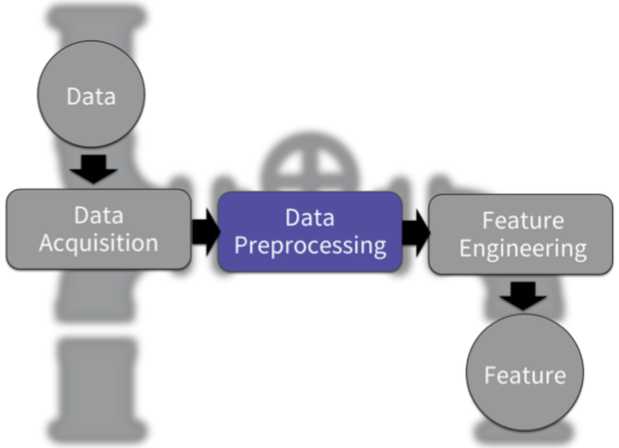
도메인에 많은 영향을 받습니다 (**Domain Specific**)



Data Preprocessing

컴퓨터가 좀 더 잘 받아들일 수 있는 형태로

Data를 가공하는 작업입니다



Techniques of Data Preprocessing

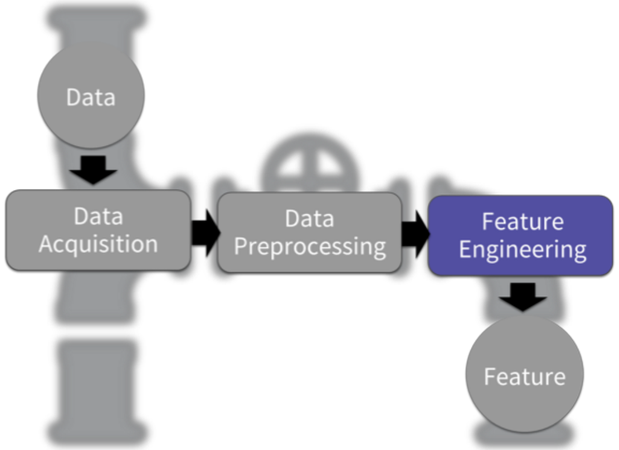
- Vectorization
- Normalization
- Handling Missing Values

Feature Engineering

도메인 지식을 활용하여

머신러닝 알고리즘이 학습을 잘 진행할 수 있도록

Preprocessed Data를 **변환**하는 작업



Techniques of Feature Engineering

- Feature Transformation
- Feature Generation
- Feature Selection
- Feature Extraction
- ……

Types of Feature

Numerical

Age, Height, Price

Categorical

Gender, Class, Job