Chapter1.
The Machine Learning Pipeline

Inju Lee

1. Data

- Observations of real-world phenomena
- 수많은 관찰값들을 통해 전체적인 모습을 짐작해볼 수 있음
- But there is always measurement noise and missing piece

2. Tasks

- 데이터를 통해 우리가 가진 문제에 대한 답을 구할 수 있음
- Workflows with data are frequently multistage, iterative processes
- 이 과정은 크게 model과 feature로 두 가지로 이루어짐

3. Models

- 데이터를 통해 전체 모습을 보고자 하지만 항상 불완전함 → Mathematical modeling
- Mathematical modeling: the relationships between different aspects of the data
- Mathematical formulas relate numeric quantities to each other
- When raw data is not numeric → feature 사용

4. Features

- Feature: a numeric representation of raw data
- Feature engineering: the process of formulating the most appropriate features given the data, the model and the task

5. Model evaluation

- 우리는 model뿐만 아니라 어떤 feature을 사용할지도 결정해야함
- Good features → modeling이 더 쉽고, 주어진 과제를 잘 완료해낼 가능성이 더 높음

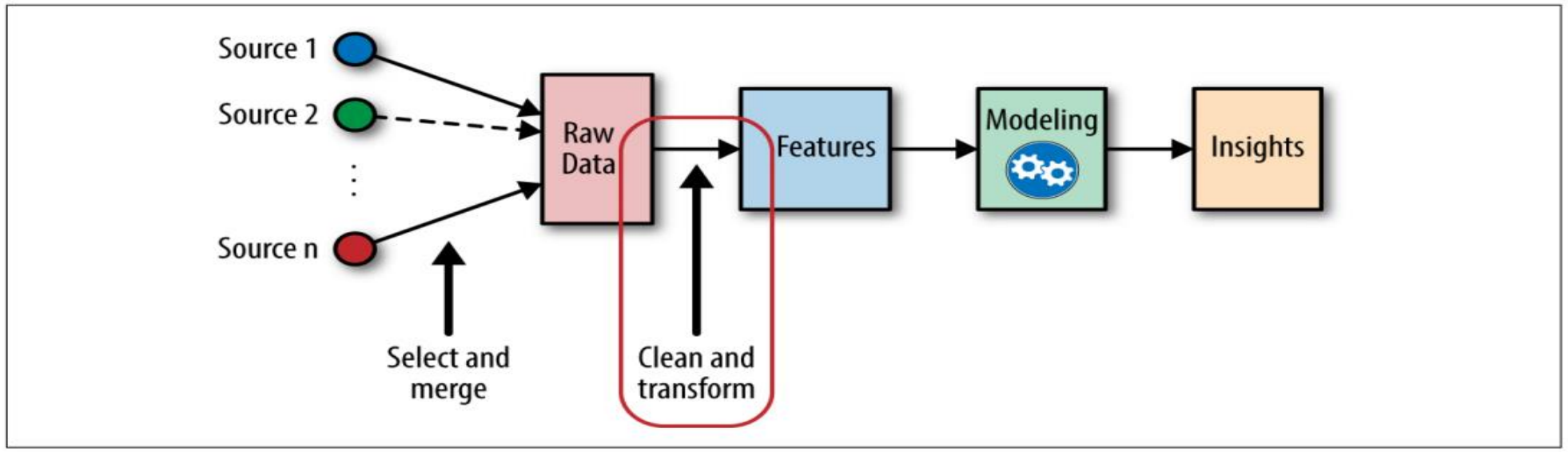

Figure 1-2. The place of feature engineering in the machine learning workflow