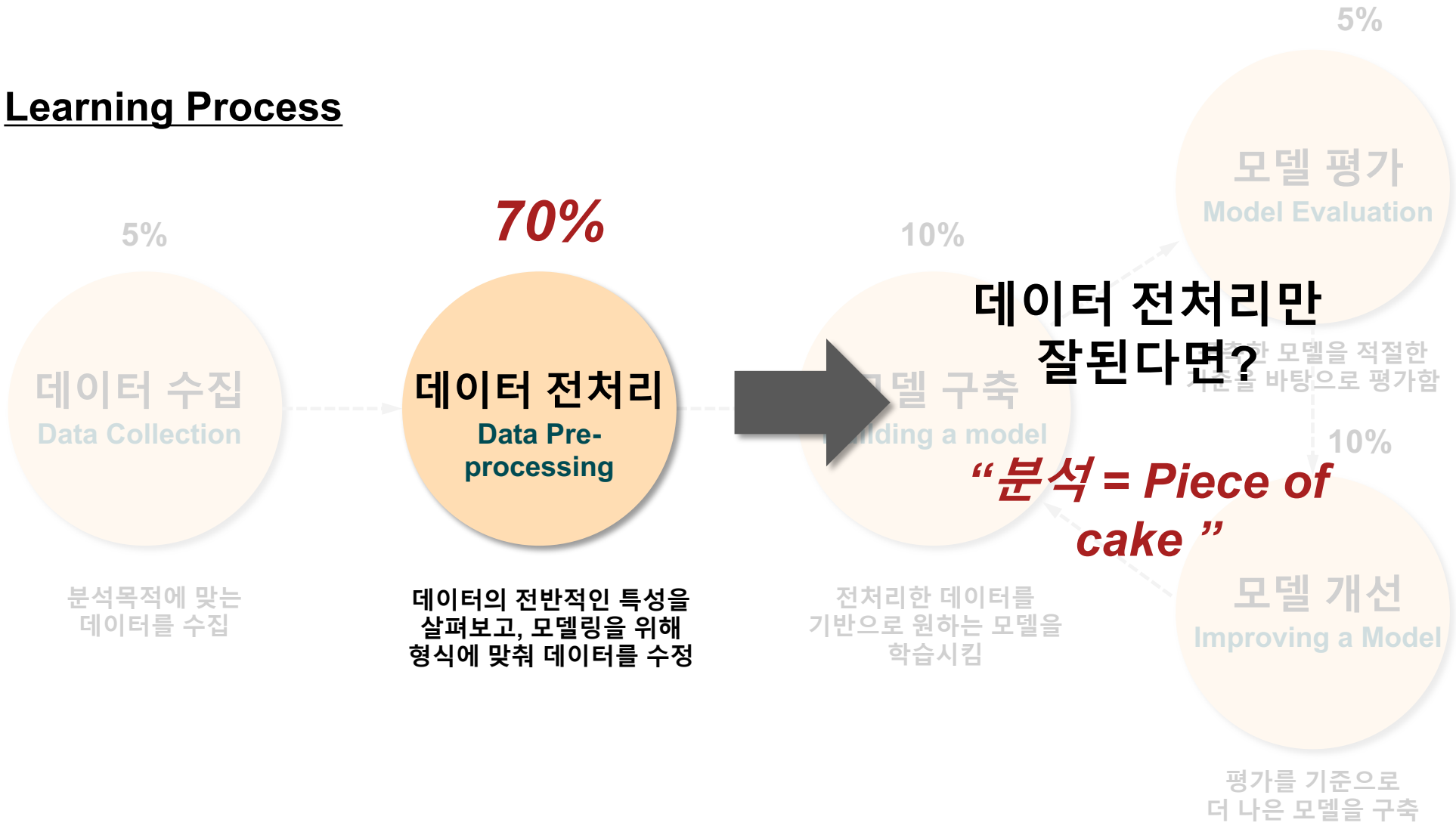
Ⅰ. 금융 빅데이터와 데이터 분석 (Big data in finance & financial time series data analysis)

❻ 데이터 전처리 과정

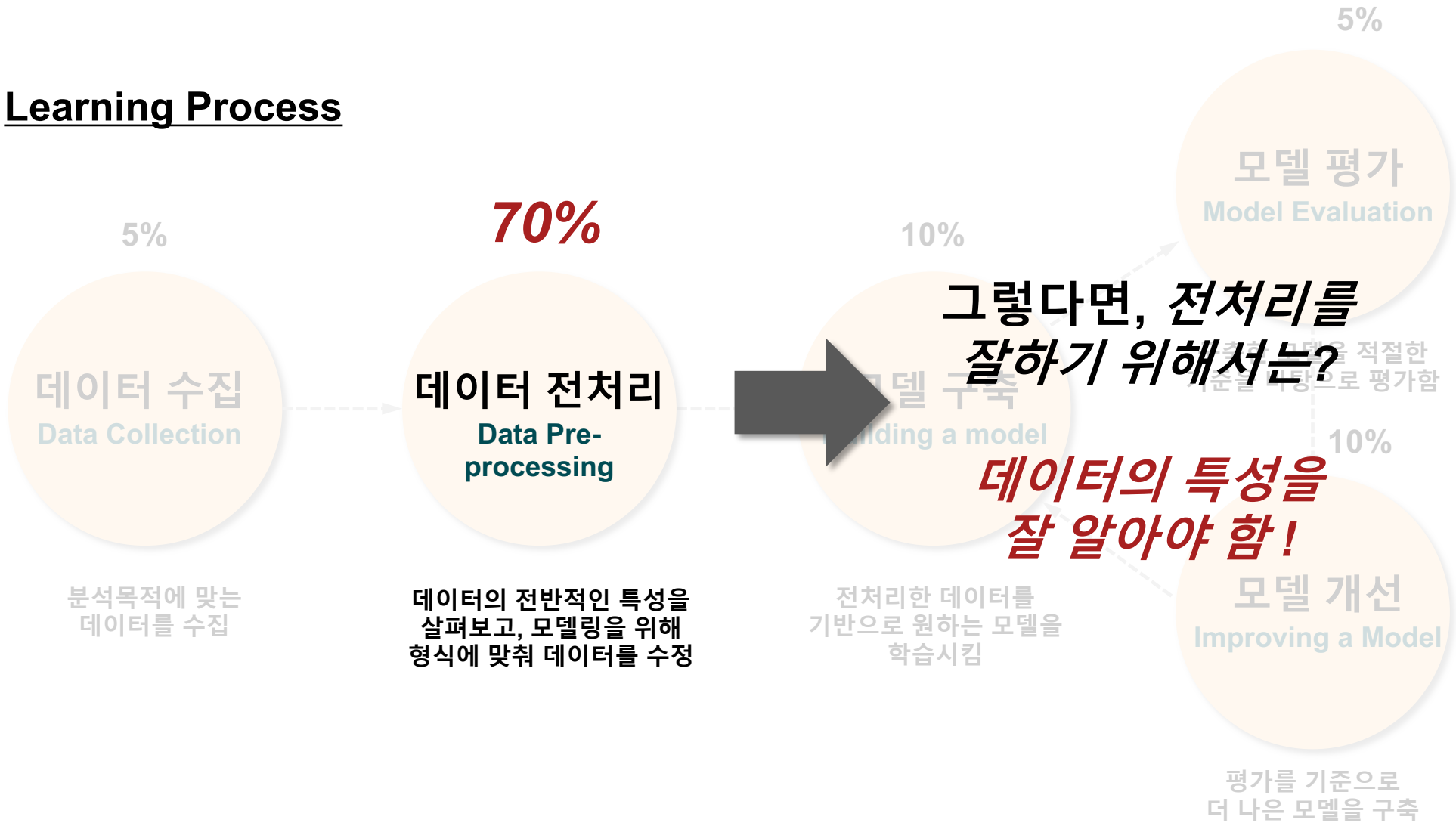
탐색적 자료분석(Exploratory Data Analysis)이란?

Learning Process



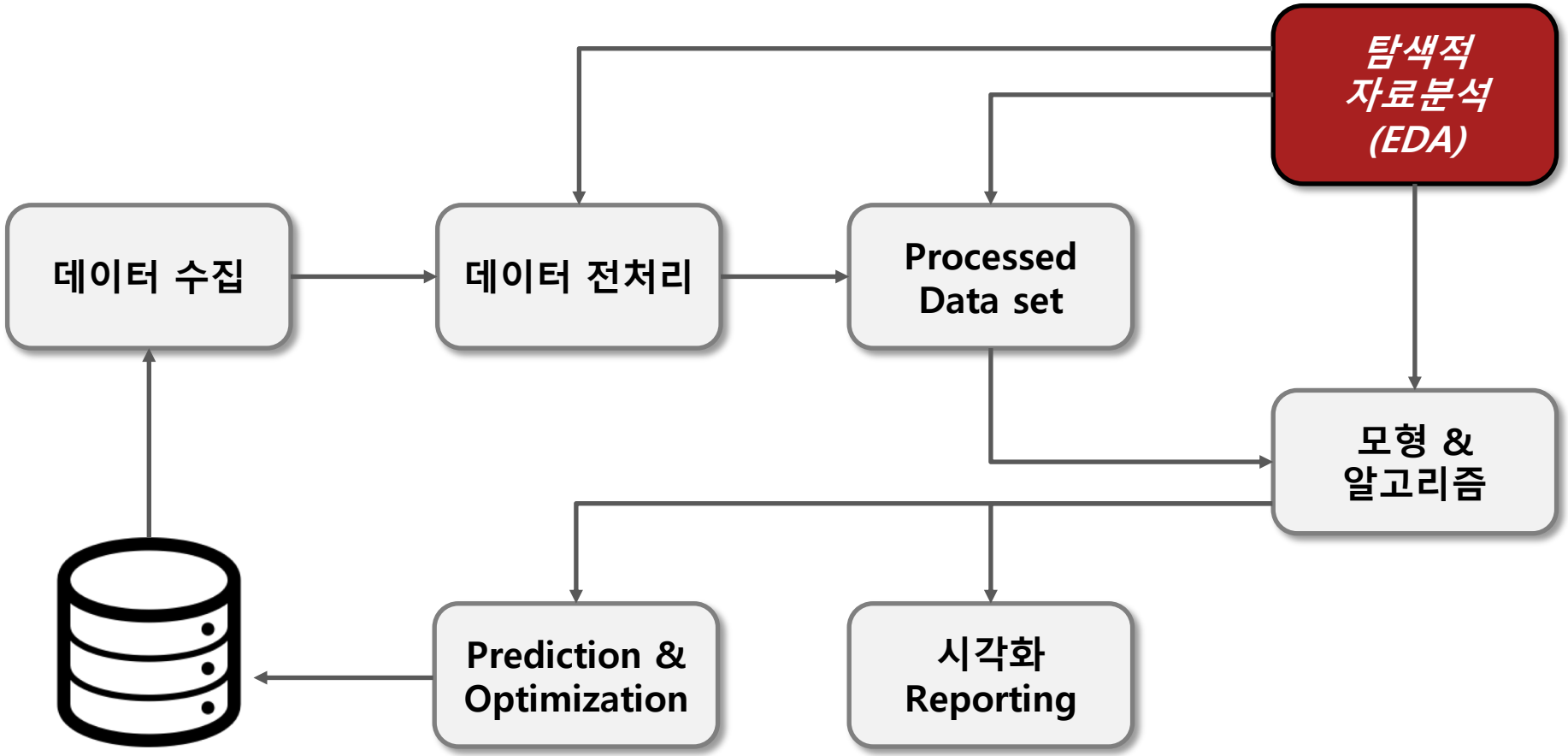
탐색적 자료분석(Exploratory Data Analysis)이란?

Learning Process

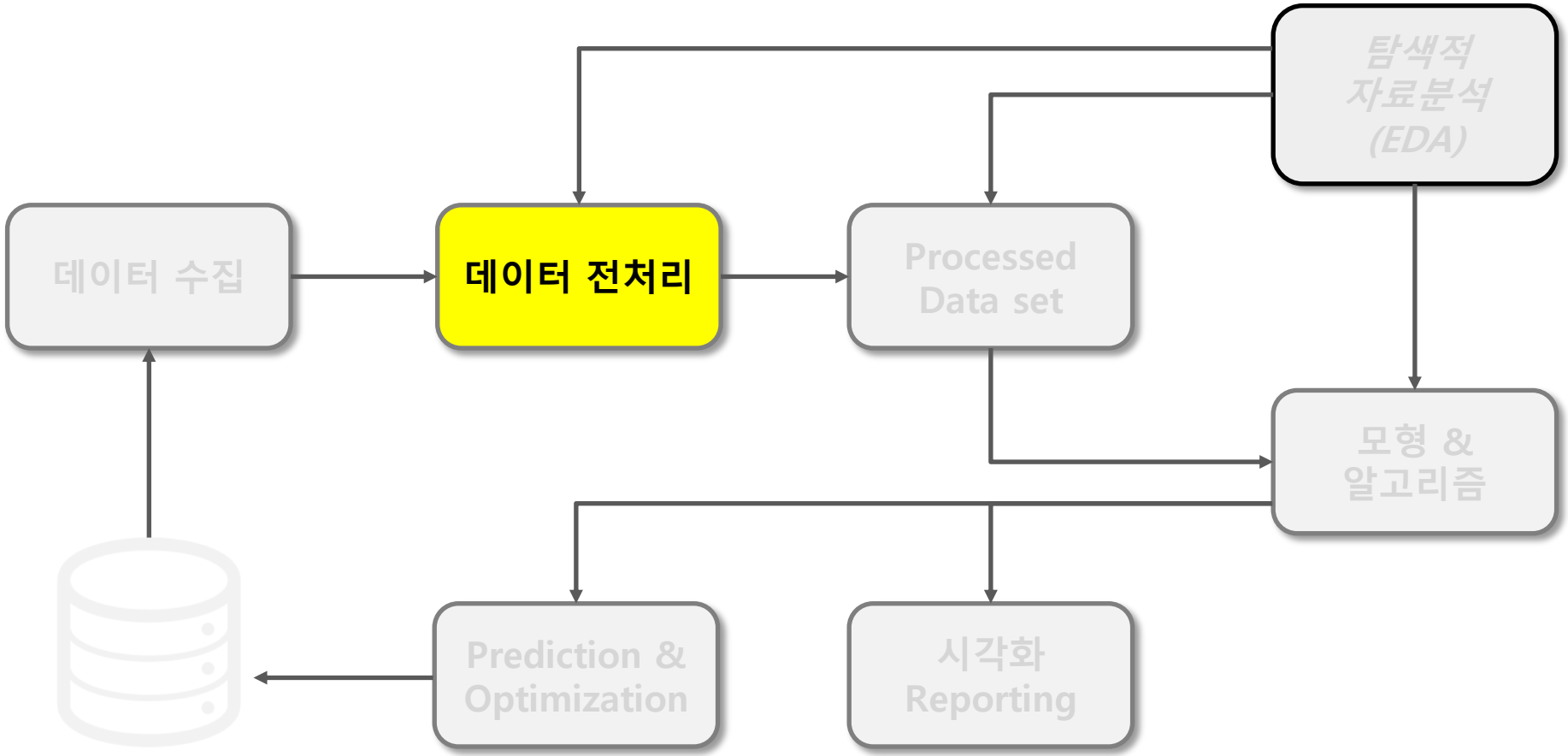


탐색적 자료분석(Exploratory Data Analysis)

모형이 얼마나 좋은 성능을 내느냐는 전적으로 데이터의 품질과 데이터에 담긴 정보량에 달려 있음.
따라서, 가능한 정보를 잃지 않으면서 모형이 잘 학습하도록 하는 것이 매우 중요함



데이터 전처리 (Data Preprocessing)



데이터 전처리 과정의 필요성

- 실생활에서 수집된 데이터는 완벽할 수 없음. 불완전성, 불일치성, 잡음 등의 현상이 항상 존재함.
- 데이터의 품질은 분석 결과에 큰 영향을 미치므로, **데이터의 전처리 과정 (Data Preprocessing)** 이 필요함. *-> 데이터 전처리 작업이 분석 작업보다 더 오래 걸릴 수 있음.*
- 데이터는 **적시성 (Timeliness)**과 결과와의 **관련성 (Relevance)**이 있어야 함. 결과와 시기적으로 맞지 않는 데이터는 의미가 전혀 없는 것은 아니지만, 분석 시기와 수집 시기와의 차이를 고려하여 적절히 보정되어야 함 (적시성). 또한, 수집된 데이터는 분석 결과에 필요한 정보를 충분히 가지고 있어야 함 (관련성).

데이터 전처리 과정의 필요성

- 데이터 전처리 방법으로 **정형화된 방법은 없으며** 데이터의 특성과 분석 목적에 따라 분석자의 노하우를 통해 결정됨.
  - *-> 시행 착오, Cross validation 등을 통해 결과의 정확도가 높아지도록 전처리 방법을 연구해야 함*
- 분석자는 데이터를 분석하는 기법 뿐만 아니라 **데이터 자체에 대한 지식이 풍부해야 함.** 의미 있는 분석을 위해서는 해당 분야 (해당 비지니스) 의 데이터를 충분히 이해하는 데이터 전문가가 필요함
  - *-> 데이터 사이언티스트의 필요성. (예 : 금융 데이터 분석을 위해서는 데이터 분석 기법 뿐만 아니라 금융 이론에 대한 깊은 이해가 필요함.)*

데이터 전처리 과정의 대표적 유형

- **Data Cleaning (정제) :** 데이터 누락이나, 중복, 잡음 , 이상치 (Outlier) 들을 수정 하거나 제거해야 함.
- **Data Integration (통합) :** 데이터는 다양한 소스로부터 수집할 필요가 있음.
- **Data Transformation (변환) :** 데이터는 속성 별로 스케일이나 단위 등이 다를 수 있으므로, 이를 표준화 할 필요가 있음. (ex: Z-Score or Min-max Normalization)
- **Data Reduction (축소) :** 분석 결과에 영향이 없다면, 데이터의 차원을 줄이는 것이 필요함. 중요도 분석을 통해 중요한 데이터에 더욱 집중해야 함 (차원의 저주).
- **Data Discretization (이산화) :** 연속된 수치 데이터를 이산화 시킴. 이산화된 숫자, 분류 명이나 단위 명 등으로 대체함.

데이터 정제 (Data Cleaning)

- **누락 데이터 처리 방안 (예시)**
  - 평균, 중앙값, 최빈수 등으로 중심 경향 적인 데이터로 채우는 방법 : 단순한 방법이면서 크게 무리가 없는 방식이나, 데이터의 분산이 감소하고, 소수의 평균이 전체를 대표하는 경우가 발생하거나, 극단적인 값으로 평균이 영향을 받음.
  - 데이터 분포에 맞게 랜덤 값을 생성하여 채우는 방법 : 데이터의 분포를 알고 있는 경우에만 적용 가능함.
  - 회귀 분석에 의한 추정 값으로 채우는 방법 : 데이터 특성에 타당한 회귀식으로 중간의 누락 값을 추정함.

데이터 정제 (Data Cleaning)

- **이상 데이터 검출 (예시)**
  - 분석자가 Min, Max 범위를 설정하고, 이 범위를 벗어난 데이터들을 이상치로 판단함.
  - 데이터의 평균과 표준편차를 계산하고 평균에서 너무 떨어진 데이터들을 이상치로 판단함.
    - (예 : ± 3σ 를 벗어남)
  - 데이터 간 최 근접 거리 (Nearest neighbor) 를 계산하여, 거리의 분포에서 멀리 있는 데이터를 이상치로 판단함.
  - 데이터를 여러 개의 클러스터로 묶어 보고 (군집화, 그룹화), 각 그룹의 중심에서 멀리 떨어진 데이터를 이상치로 판단.
  - 이상치도 하나의 정보가 될 수도 있음.

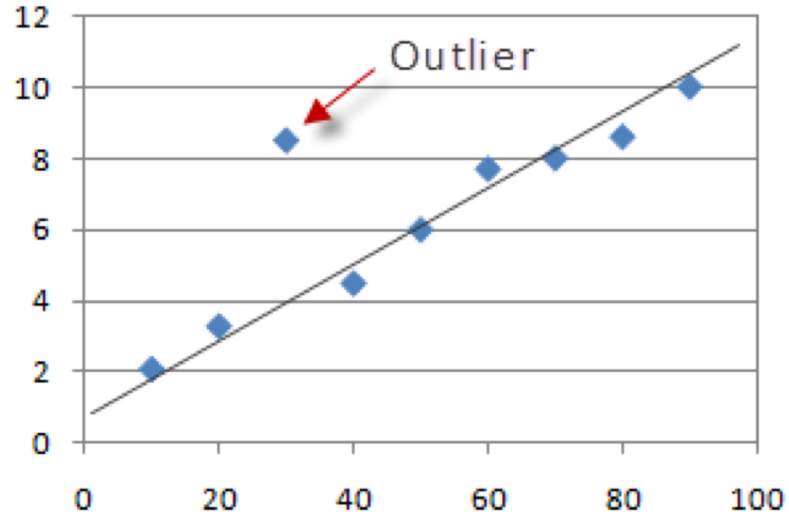
데이터 정제 (Data Cleaning)

- **잡음 데이터 검출 (예시)**
  - 이동 평균에 의한 필터링: 이동 평균에 비해 너무 높거나 낮은 값들을 잡음으로 판단함. 이동 평균으로 대체.
  - Kernel 이나 Spline 등의 평활화에 의한 필터링 : 이동 평균은 데이터 후행성이 나타나므로, Kernel이나 Spline 사용.
  - 푸리에 변환 등을 통한 Low pass, High pass 필터링.

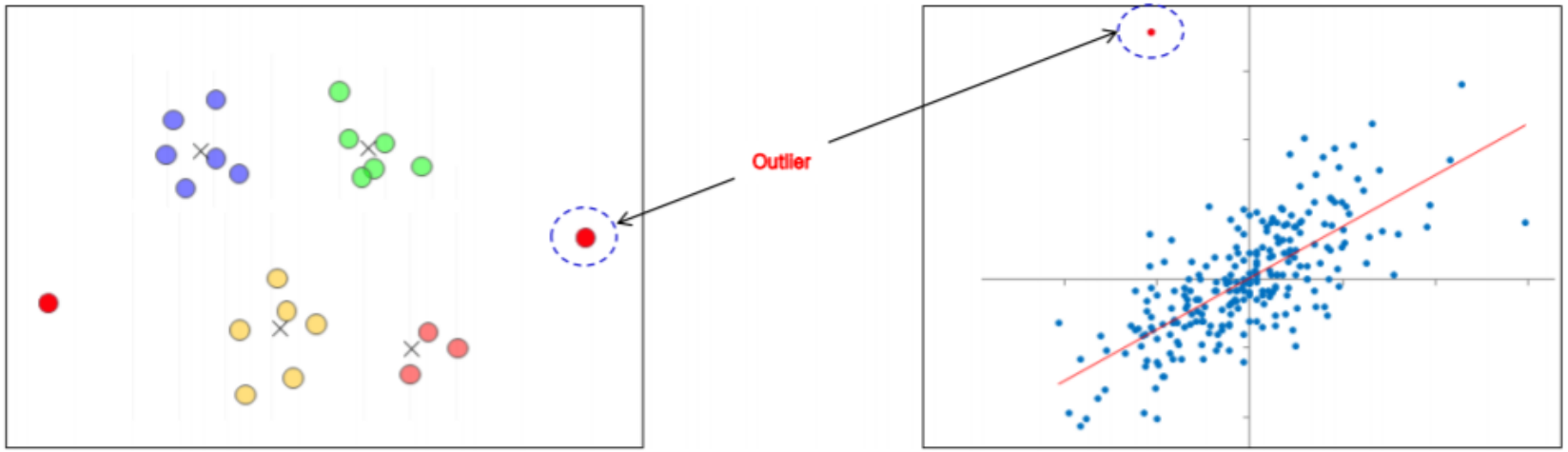
데이터 변환 (Data Transformation)

- **함수 식에 의한 데이터 변환 (예시)**
  - 데이터의 범위가 넓고 수치가 낮은 쪽이나 높은 쪽으로 크게 치우친 경우 로그나 제곱근 함수로 변환이 필요한 경우도 있음 (Right or Left Skewness). 예를 들어 장기간 주가 데이터의 경우 현재 주가가 과거 주가에 비해 높은 경우, 과거 주가의 등락은 잘 드러나지 않으므로 절대적 수준 보다 등락의 변화를 분석하고자 할 경우는 로그 주가를 이용함.
  - 함수 식을 이용하여 데이터를 변환하는 경우는 데이터의 왜곡이 발생하므로 결과 해석에 이를 반영해야 함.
  - 데이터 간 비교를 위해서는 **평균과 분산을 표준화** 시킬 필요가 있음. (Z-score Normalization)
  - 너무 큰 값이나, 작은 값이 결과에 미치는 영향을 줄이기 위해 데이터를 일정 범위로 맵핑 시킬 필요가 있음. (예 : 직선의 방정식이나 시그모이드 같은 함수를 이용하여 데이터의 범위가 0 ~ 1 또는 -1 ~ +1 사이에 있도록 변환)

데이터 이상치 (Outlier) 유형 (예시)



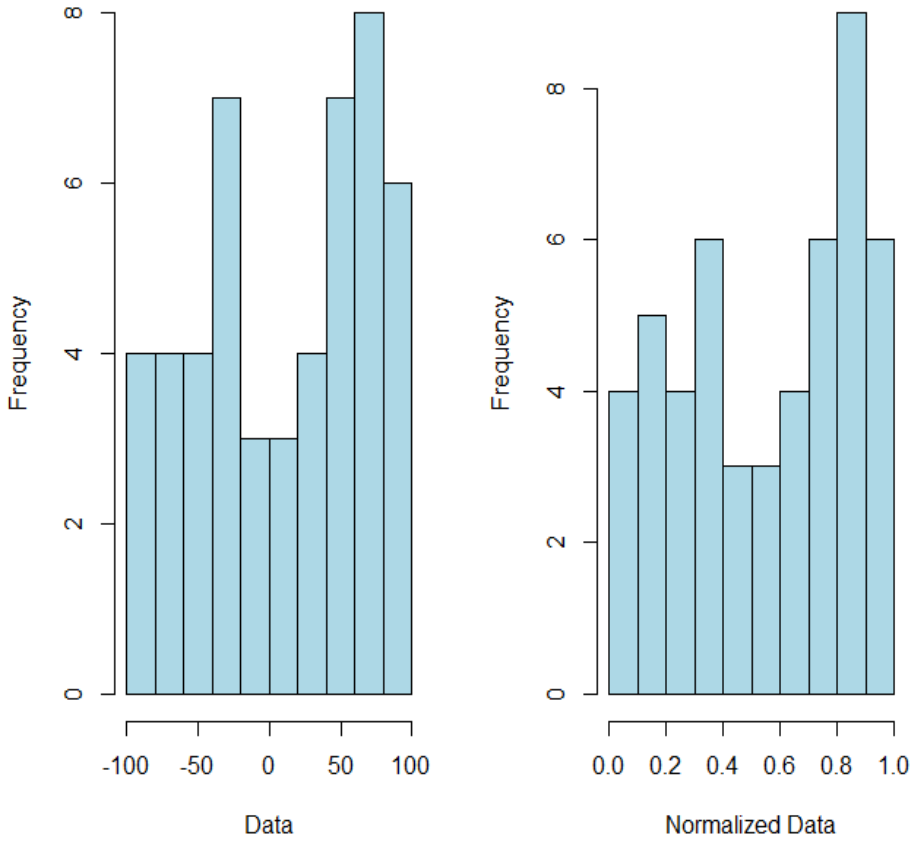
▪ 데이터 이상치 (Outlier) 유형 (예시)



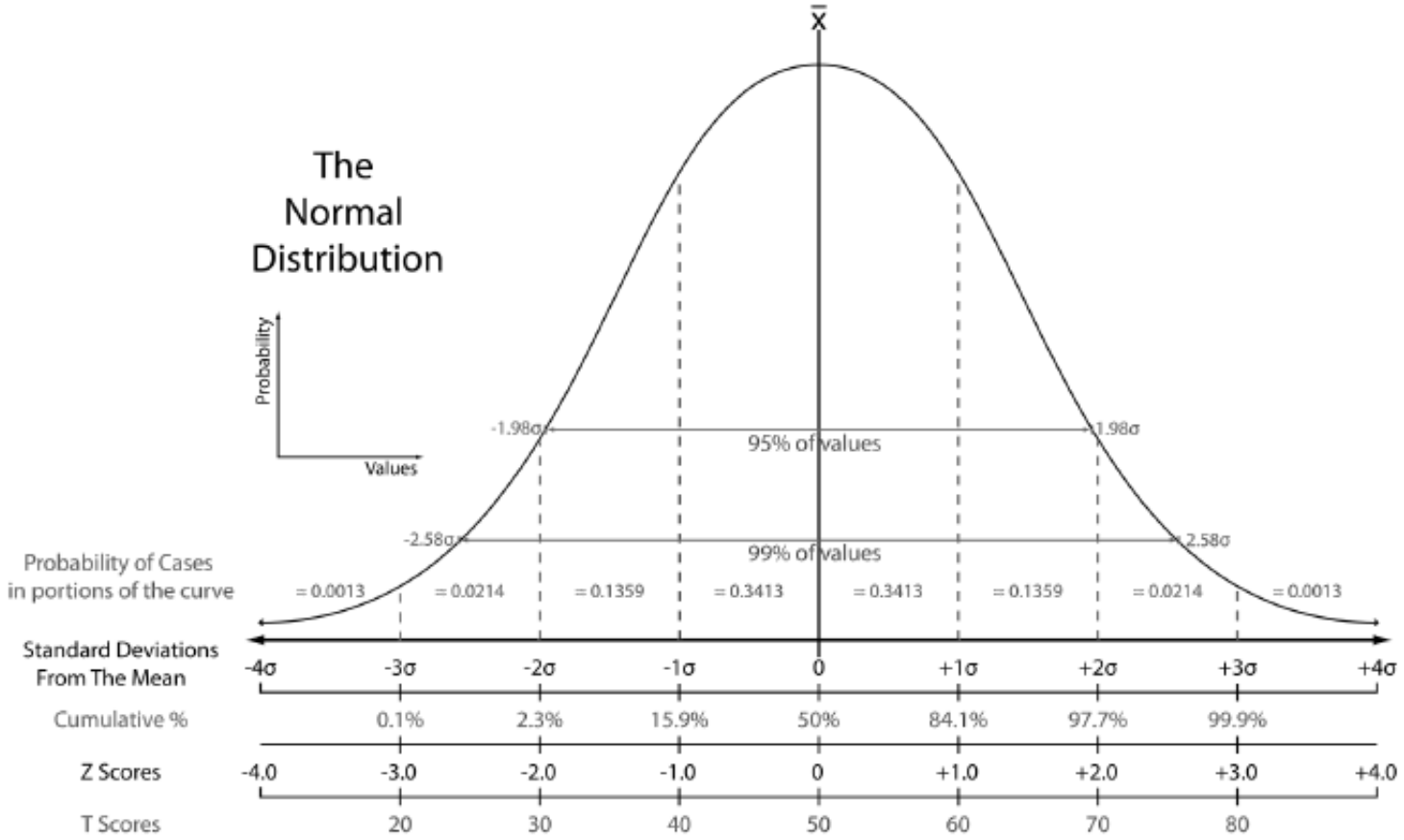
데이터 변환 (Data Transformation) - 표준화 (Normalization)

- 데이터 분석을 위해서는 데이터를 표준화 시켜야 할 경우가 많음. 데이터의 크기가 상이할 경우 분석 결과에 영향을 미칠 수 있으므로 스케일을 맞추어야 함.
- 데이터 표준화 방법은 **Z-score normalization, Min-max normalization** 등이 있음

Z-score Normalization → $x = \frac{(x - \overline{x})}{\sigma}$



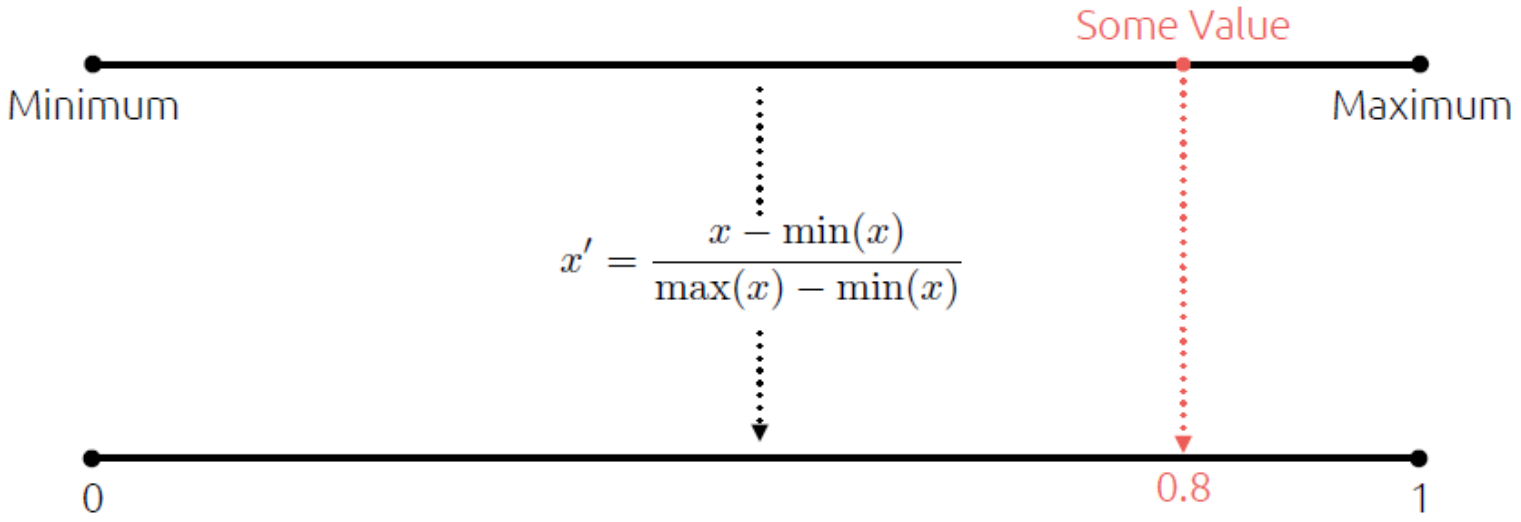
Standardization(표준화) : Z-scoring



" -∞에서 +∞까지의 값을 지님 "

※ 참고 – Normalization

Min-Max Normalization(정규화) : Rescaling



" 0에서 1사이의 값을 지님 "

Ⅰ. 금융 빅데이터와 데이터 분석
(Big data in finance & financial time series data analysis)

❼ 텍스트 마이닝(Text Mining)

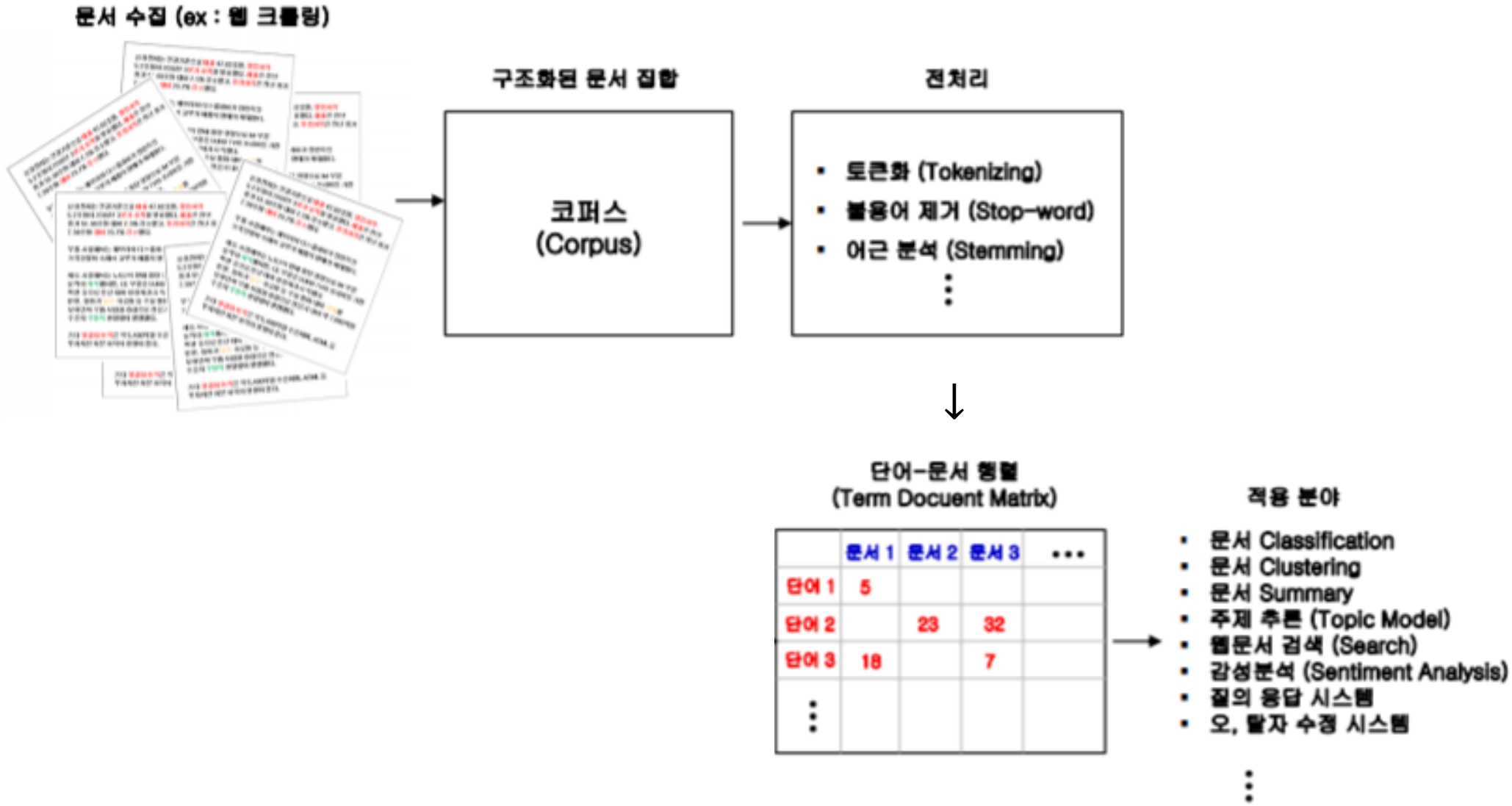
텍스트마이닝(Text Mining)이란?

- 텍스트 마이닝은 텍스트로 구성된 다수의 문서에서 특징들을 (Features) 추출하여 의미 있는 정보를 추론하는 분야임. (문서 분류, 주제 추론 등)
- 텍스트 마이닝은 통계학과 기계학습 기술 뿐만 아니라 언어학 분야까지 (파싱, 형태소, 구문, 어휘, 품사 분석 등) 포함한 **자연어 처리 기술 (NLP)을 기반**으로 함.
- 텍스트 문서는 문장, 어절, 어구, 단어, 형태소로 분해되어 기계가 이해할 수 있는 형태로 구조화 됨. **-> 학습 데이터**
- 문서의 내용들은 최소 분석 단위로 분해되고 불필요한 요소들을 제거하거나 대체하는 등 정제 과정 (Cleaning)을 거침. **-> 전처리 (Preprocessing)**

텍스트마이닝(Text Mining)이란?

- 문서의 내용은 인간이 이해하는 것처럼 언어적으로 분석하기보다는 문서에 사용된 단어들의 빈도수에 기반하여 통계적으로 분석함.
- 텍스트로 구성된 데이터는 (비정형 데이터) 통계분석을 위해 수치 데이터 (정형화된 데이터)로 변환됨. (예 : Term Document Matrix, tf-idf 등)
- 텍스트 마이닝은 문서 작성자의 의도는 모르더라도 단어들의 발생 빈도 등을 분석하여 숨겨진 관계와 경향 등을 발견함.

Text Mining(텍스트마이닝) 분석 절차



텍스트 마이닝 용어

- **Corpus** : 텍스트 문서의 집합체. 비정형화된 문서를 1차적으로 정형화함. (예 : 문서 작성일, 작성자, 내용 등)
- **Unigram** : 개별 단어. 각 단어들은 서로 독립적임.
- **Bigram** : 두 개의 연속된 단어. 어떤 단어 뒤에 어떤 단어가 나올 확률이 높은지, 어떤 단어는 나올 수 없는지 등을 알 수 있음. (종속적, 1차 마코프 연쇄 과정)
- **N-gram** : N-개의 연속된 단어. 단어들의 선, 후 관계를 파악할 수 있으나 조합이 너무 많아 N을 크게 하기 어려움.
- **Noun phrases** : 명사구. 단어만으로 분석하는 것보다 가능한 구 (Phrases) 단위로 분석하면 더 유용한 정보를 얻을 수 있음. (예 : "파란 차", "내가 사는 곳" 등)
- **Token (토큰)** : 의미를 가진 최소 단위. (기호, 단어, 구 등). 텍스트 분석의 최소 단위.
- **Stop-words removal (불용어 제거)** : 분석이 불필요한 토큰을 제거함. (예 : is, am, it 등). 그러나 "To be or not to be" 와 같이 필요한 경우도 있을 수 있음.

텍스트 마이닝 용어

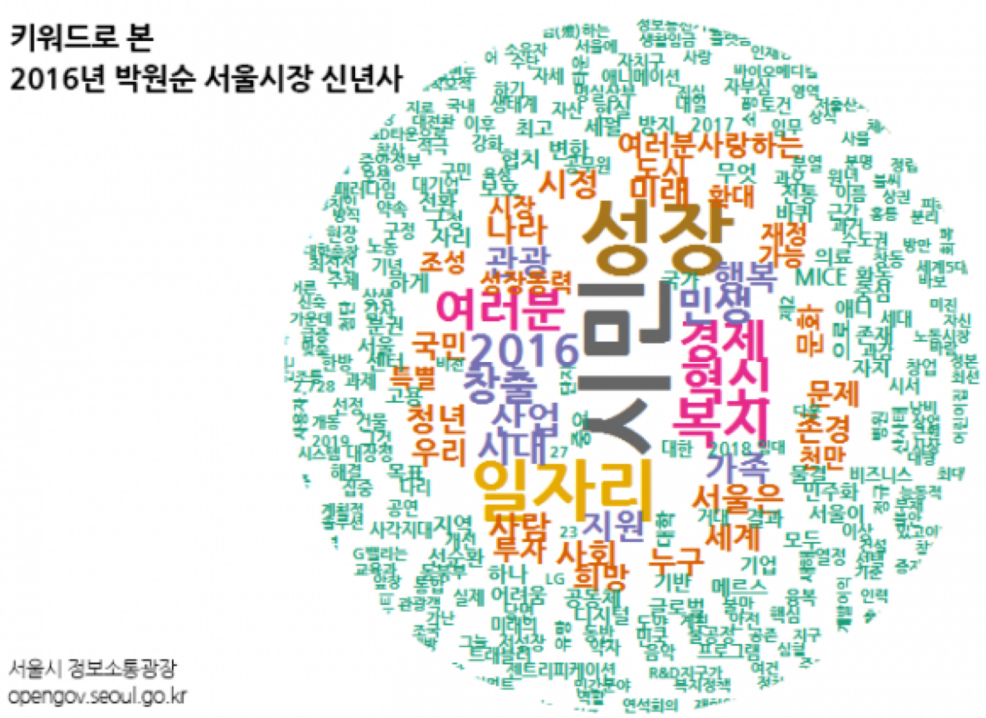
- **Stemming (어근 분석)** : 기본 단어에서 파생된 단어는 기본 단어로 대체함. (예 : making, makes, made, make, 동의어 압축 등).
- **Term-Document matrix (TDM, 단어-문서 행렬)** : 문서별로 사용된 단어의 빈도를 행렬 형태로 표시함.
- **Bag of words (단어 집합)** : 여러 문서에 등장하는 단어들을 모아 놓은 집합체.
- **Term frequency (TF)** : 어떤 문서에 등장하는 단어의 빈도수. TF가 높으면 이 문서에서 이 단어는 중요하다고 판단함.
- **Document frequency (DF)** : 어떤 단어가 얼마나 많은 문서에 등장하는지에 대한 지수. DF가 높을수록 범용적인 단어이므로 중요도가 낮음.
- **Inverse document frequency (IDF)** : DF의 역수. IDF가 높을수록 중요도가 높음.
- **TF-IDF** : TF와 IDF의 곱. 단어의 중요도를 나타내는 척도. -> 단어의 중요도 지수
- **Similarity Measure (문서의 유사도)** : 문서 간의 유사도를 평가함. TDM 공간에서 TF-IDF 간의 (코사인) 거리를 측정함. -> Vector Space Model

데이터 마이닝과 텍스트 마이닝의 차이점

	데이터 마이닝	텍스트 마이닝
대상	수치 또는 범주화된 데이터	텍스트
구조	관계형 데이터 구조	비정형 또는 정형의 텍스트 데이터
목적	미래 상황 결과의 예견 및 예측함	적합한 정보를 획득하고 의미를 정제하고 범주화 함
방법	기계학습	기계학습, 인덱싱, 언어 처리, 온톨로지(Ontology) 등

텍스트 마이닝 (Text Mining) 예시

빅 데이터로 본 경북의 키워드



빅 데이터로 본 2016년
박원순 서울시장 신년사 키워드