데이터 분석 절차

고려대학교 석 준 희

목차

- 학습 이론 (Learning Theory)
- 데이터 분석 절차
- 최근접 이웃 기법 (KNN: K-Nearest Neighbor)

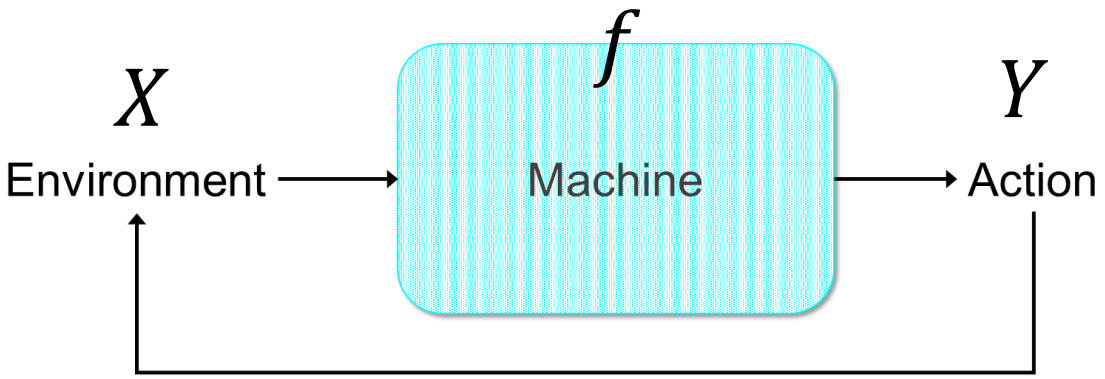
KOREA UNIVERSITY

데이터 분석 절차

학습 이론

인공지능 에이전트 (AI Agent)

- 인공지능 에이전트: 관측된 주변 환경 데이터로부터 적절한 행동에 대한 판단을 내리는 주체
  - 환경(Environment, X): 관측된 데이터, e.g.) 센서로부터 수집된 주행 환경
  - 행동(Action, Y): 판단의 결과, e.g) 가속, 감속, 핸들 조작 등
  - 에이전트 (Agent, f): X로부터 Y를 도출하는 함수로 표현 가능



$$Y = f(X)$$

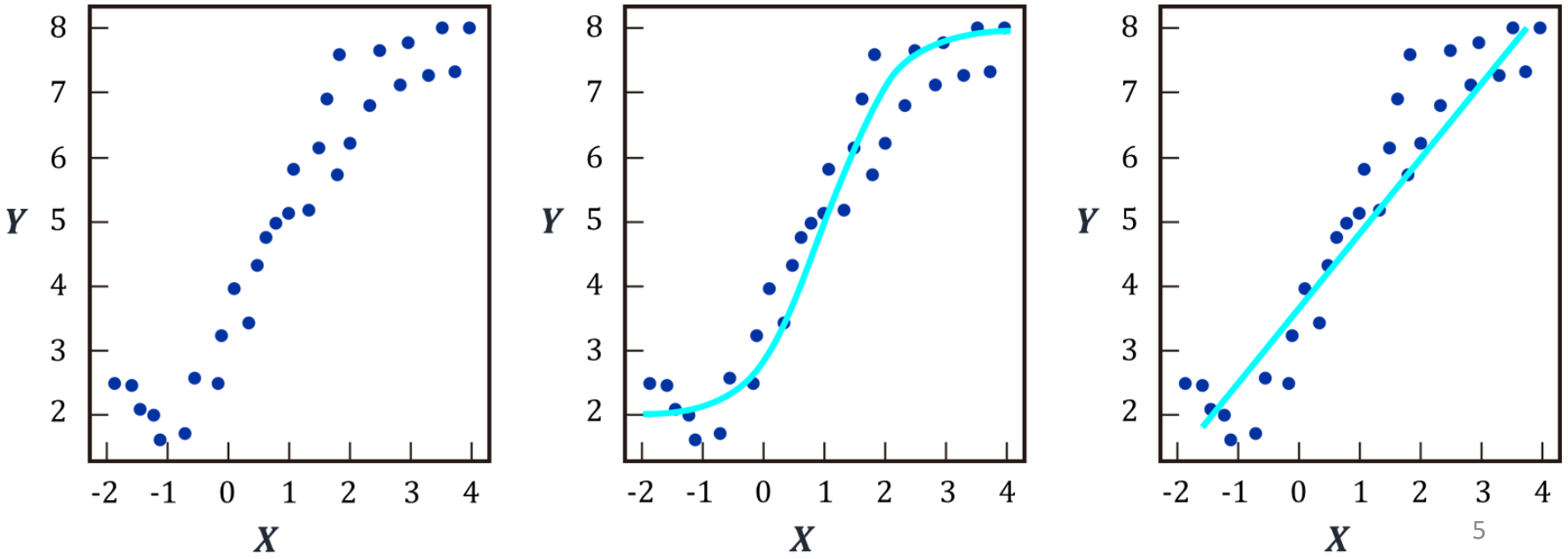
- 인공지능 에이전트의 학습 (Learning)
  - 주어진 데이터: 특정 환경 X에 대한 적절한 행동 Y의 쌍
  - 학습: 주어진 데이터로부터 X와 Y의 관계에 대한 적절한 모델 $f()$를 찾는 과정
  - 데이터기반의 인공지능 구현 = 기계학습

학습의 목표

- 출력변수 (종속변수, 관심변수) Y 와 입력변수 (독립변수, 인자) $X = (X_1, X_2, \ldots, X_p)$에 대하여, 둘 사이의 관계는 일반적으로 아래와 같이 표현 가능
  - ϵ 잔차 (Residual): 정보의 불완전성으로 모델로 설명되지 않고 남은 부분

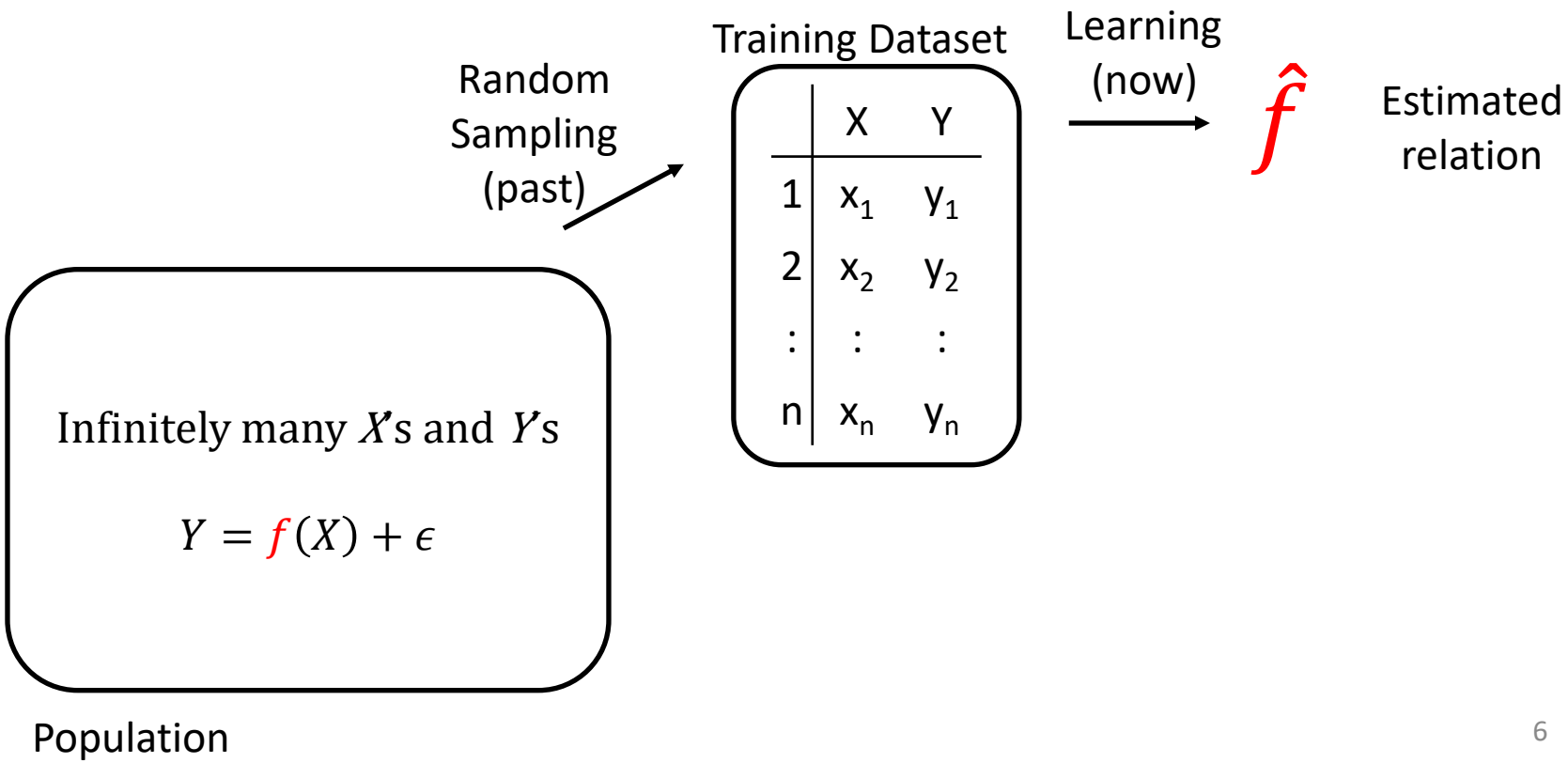
$$Y = f(X) + \epsilon$$

- 학습의 목표: $f()$ 를 주어진 관측 데이터(훈련 데이터, Training Data)로부터 추정
  - 회귀 문제 (Regression): Y가 수치형 변수일 때
  - 분류 문제 (Classification): Y가 범주형 변수일 때



학습의 문제

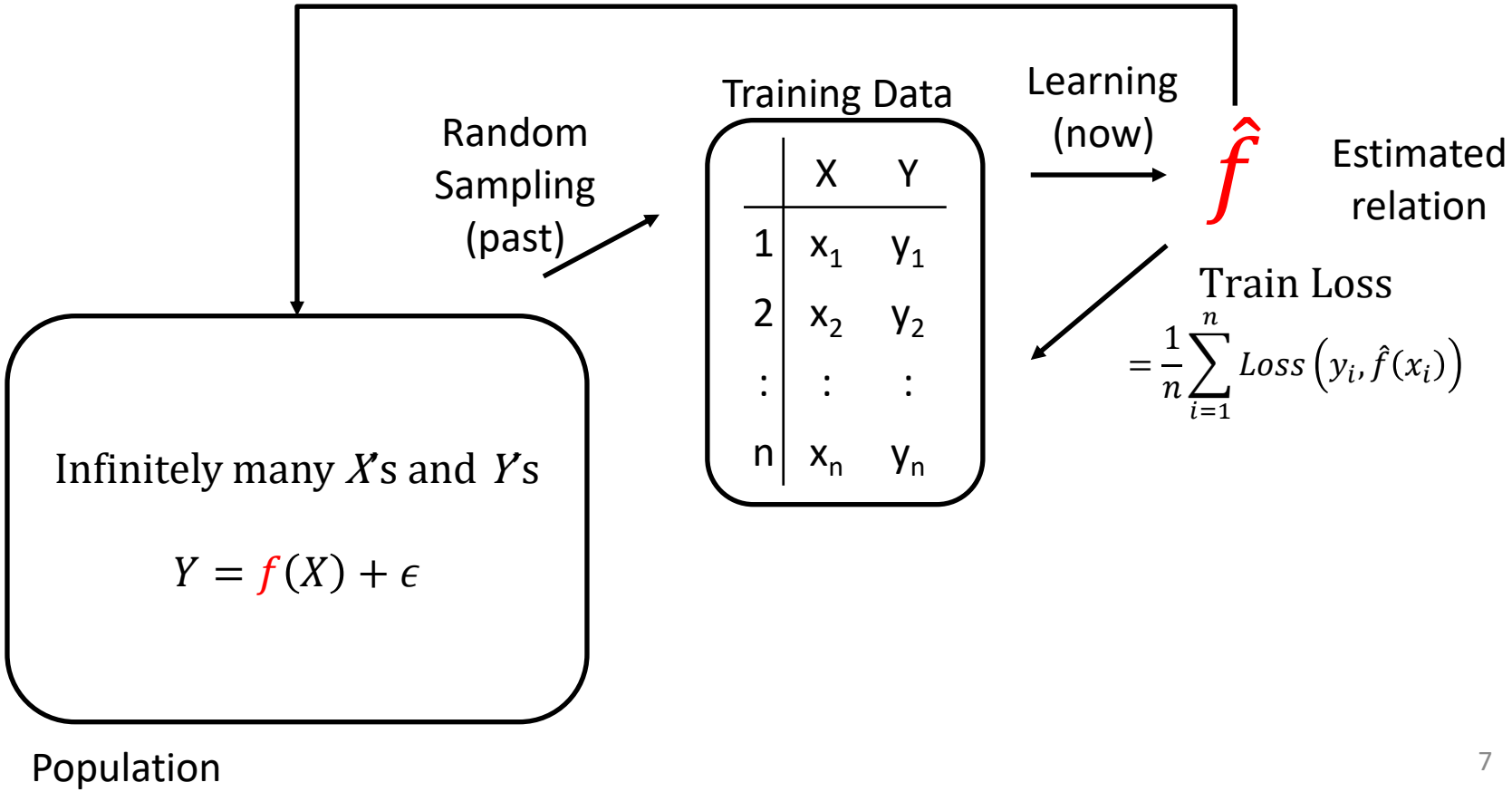
- 모집단(Population): 무한히 많은 표본을 갖는, 우리가 알고자 하는 '일반적' 대상
- 훈련 데이터셋 (Training Dataset): 관계를 추정하기 위해 관측된 데이터
- 일반적으로 어떤 손실함수(Loss Function)를 정의하고 이를 최소화하는 $\hat{f}()$ 를 선택
  - 예) 제곱에러(Square Error) 손실 $Loss\left(y, \hat{f}(x)\right) = \left(y - \hat{f}(x)\right)^2$



학습의 문제

- 학습된 모델 $\hat{f}()$을 모집단에 적용했을 때의 진짜 손실을 최소화하고 싶지만, 불가능
- 관측 가능한 훈련 손실(Train Loss)을 대신 이용할 수 있을까?

$$\text{True Loss} = \mathrm{E}\left[Loss\left(Y, \hat{f}(X)\right)\right] \cong \frac{1}{\infty}\sum_{i=1}^{\infty} Loss\left(y_i, \hat{f}(x_i)\right)$$



	X	Y
1	x_1	y_1
2	x_2	y_2
:	:	:
n	x_n	y_n

학습의 문제

- 훈련 데이터는 모델링되는 부분 $f(X)$와 잔차 부분 ϵ이 모두 관측이 가능하기 때문에, 잔차까지 모델에 포함하여 훈련 손실을 0으로 만드는 $\hat{f}()$을 찾는 것이 항상 가능

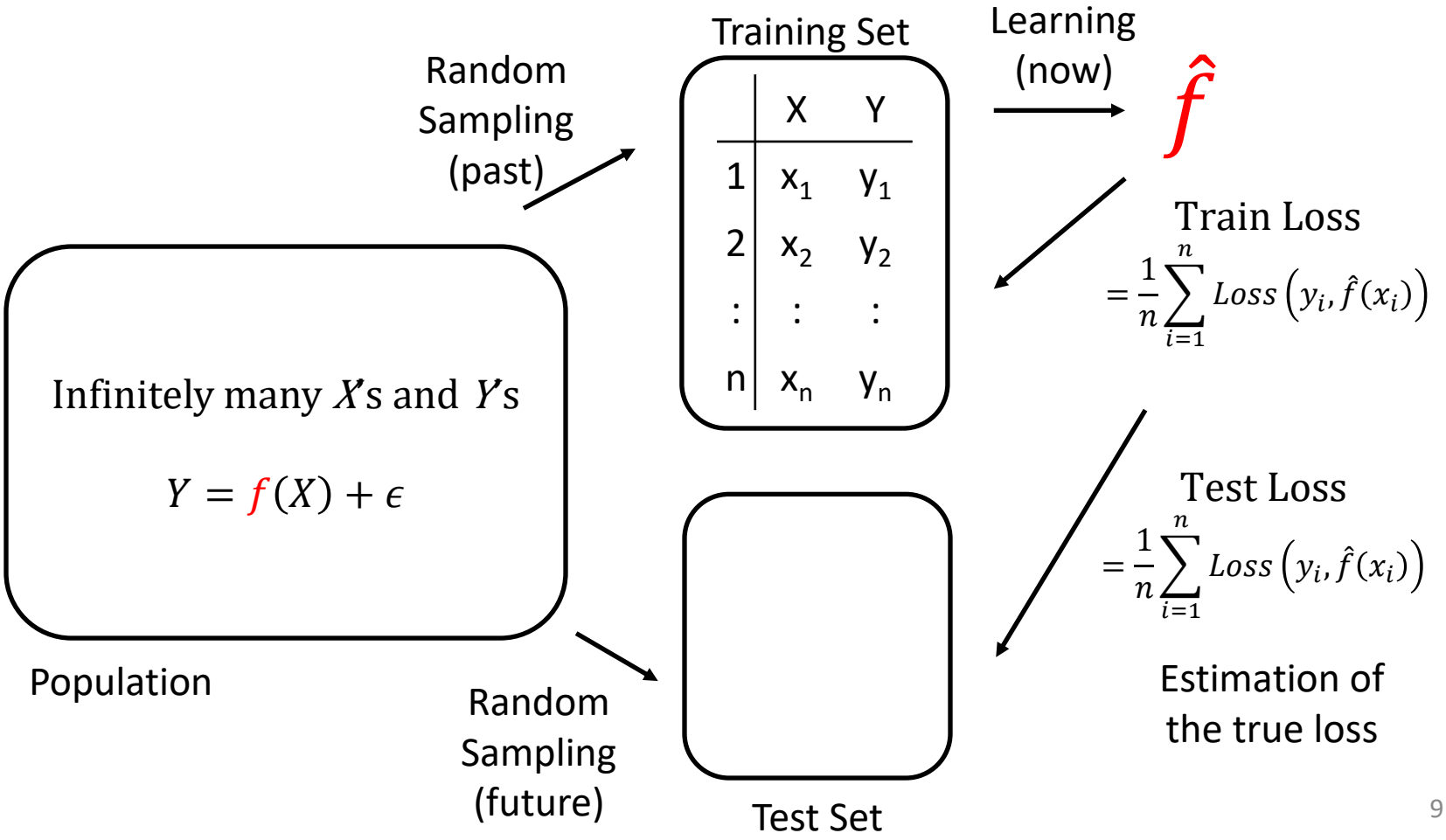
- 이는 잔차를 포함하는 모집합의 일반적인 손실을 반영하지 못하기 때문에 일반적으로 적용할 수 없음
  - 과대적합(Overfitting)의 문제: 주어진 훈련 데이터에는 낮은 손실을 보이지만, 새로운 일반적인 데이터에 대해서는 높은 손실을 보임

- 실제 손실을 추정하기 위해서는 훈련 데이터가 아니라 또 다른 데이터셋이 필요
- 평가 데이터셋 (Test Dataset)
  - 훈련 데이터와 완전히 독립적이 또다른 데이터셋으로 학습때는 모집단과 같이 관찰할 수 없음
  - 훈련과정이 종료된 후에 훈련된 모델의 실제 성능을 측정하기 위해서 사용
  - 모집단의 실제 손실을 추정할 수 있음

- 시점 상으로 과거에 수집된 훈련 데이터를 이용해 현재 모델을 훈련하고, 훈련된 모델의 성능을 미래에 수집되는 평가 데이터를 이용해 평가하는 개념

학습의 문제

- 학습의 문제: 훈련 데이터를 이용하여 평가 데이터에서 손실을 최소화하는 모델 $\hat{f}()$ 을 찾는 문제
  - 평가 데이터는 학습 시에는 관측이 불가능



학습 모델의 복잡성

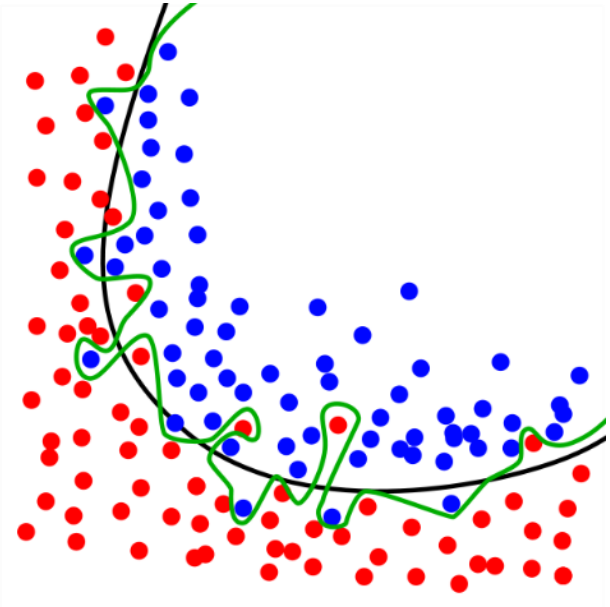
- 모델 복잡성 (Model Complexity)
  - 어떤 모델이 표현한 수 있는 데이터 패턴의 범위
  - 수치적으로 잘 정의되지는 않지만 모델들끼리의 비교를 통해 상대적으로 확인 가능
  - 어느 한 모델이 다른 모델이 표현할 수 있는 모든 패턴의 데이터를 표현할 수 있으면 더 복잡한 모델
  - 유연성(flexibility), 자유도(degree of freedom) 등으로 불리기도 함

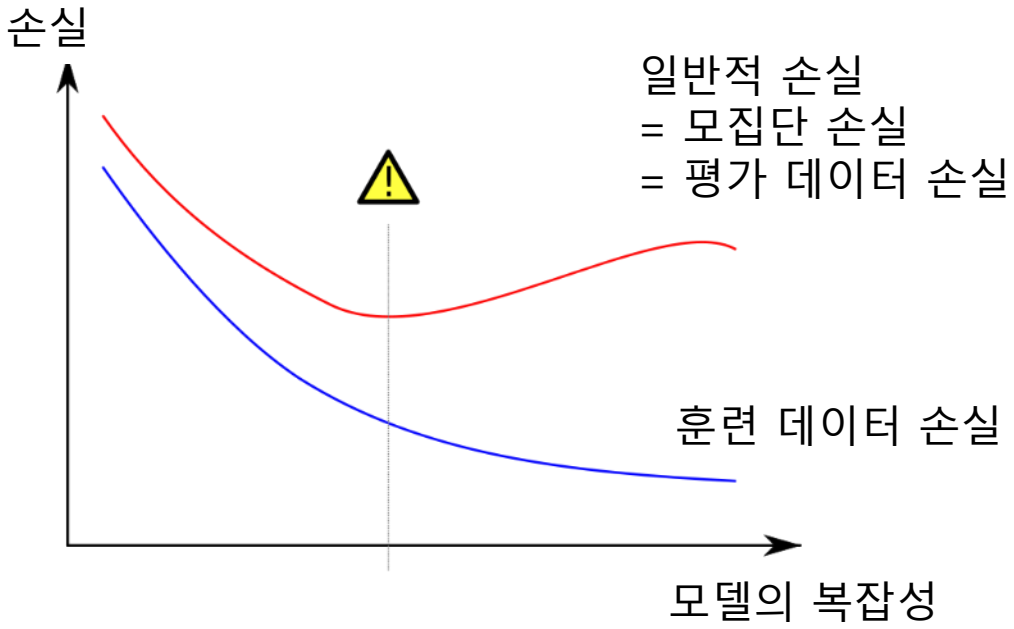
- 예시
  - $f_1(x) = ax^2 + bx + c$ vs. $f_2(x) = ax + b$
  - 선형 회귀 모델 vs. 인공신경망

- 복잡한 데이터를 학습하기 위해서는 더 복잡한 모델이 필요
- 하지만 반드시 복잡한 모델이 좋은 것은 아님!

복잡성에 따른 성능의 변화

- 모델이 복잡할 수록 훈련 데이터에서의 성능은 향상
  - 일반적인 모집단에서의 성능(평가 데이터에서 측정)은 그렇지 않음
  - 너무 복잡하지도 너무 단순하지도 않은 최적의 모델이 존재

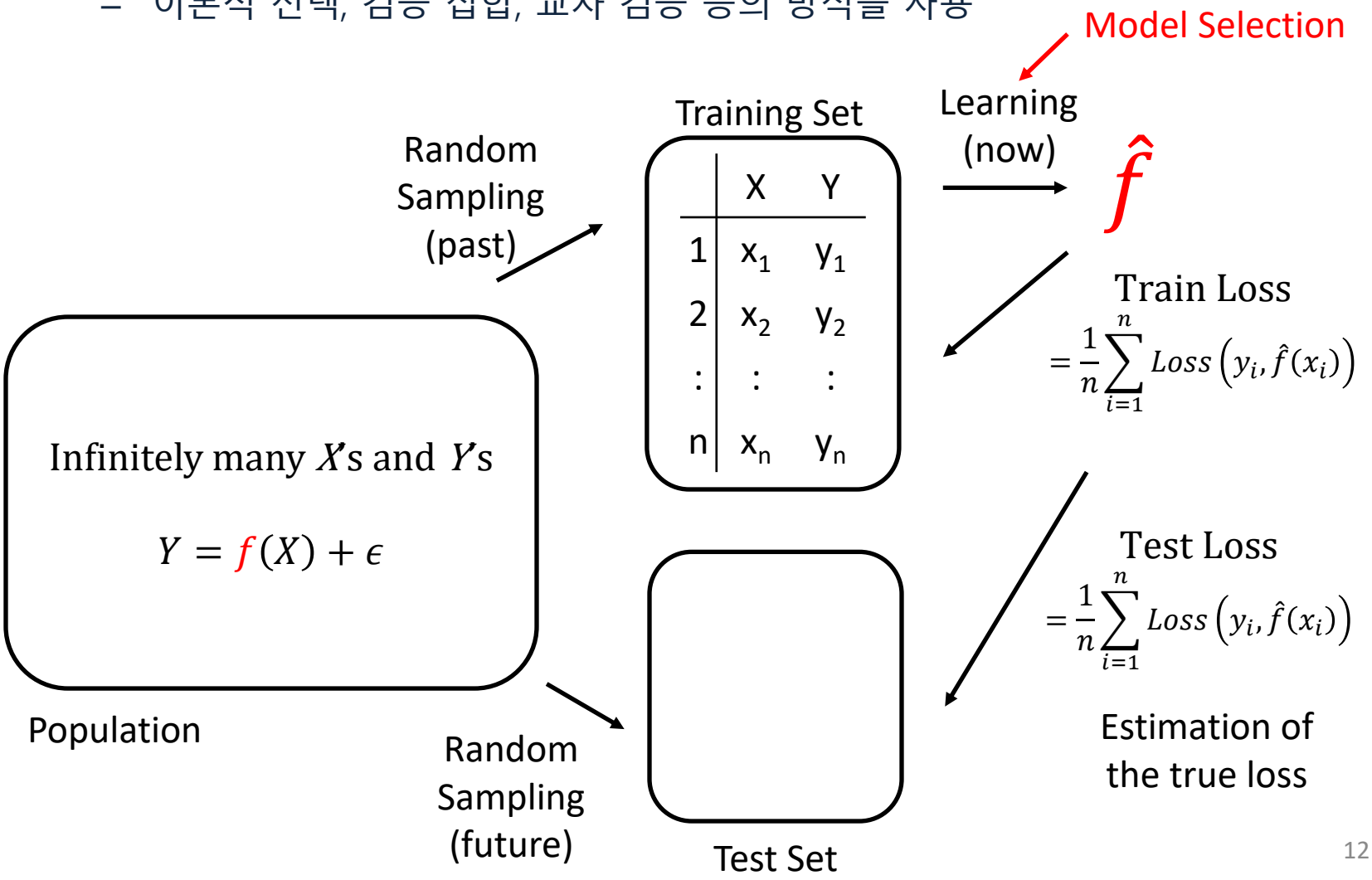




- 과소적합(Underfitting): 모델이 너무 단순해서 훈련/평가 둘 다 손실이 높은 현상
- 과대적합(Overfitting): 모델이 너무 복잡해서 훈련에서는 손실이 낮지만 평가에서는 손실이 높은 현상

모델 선택 (Model Selection)

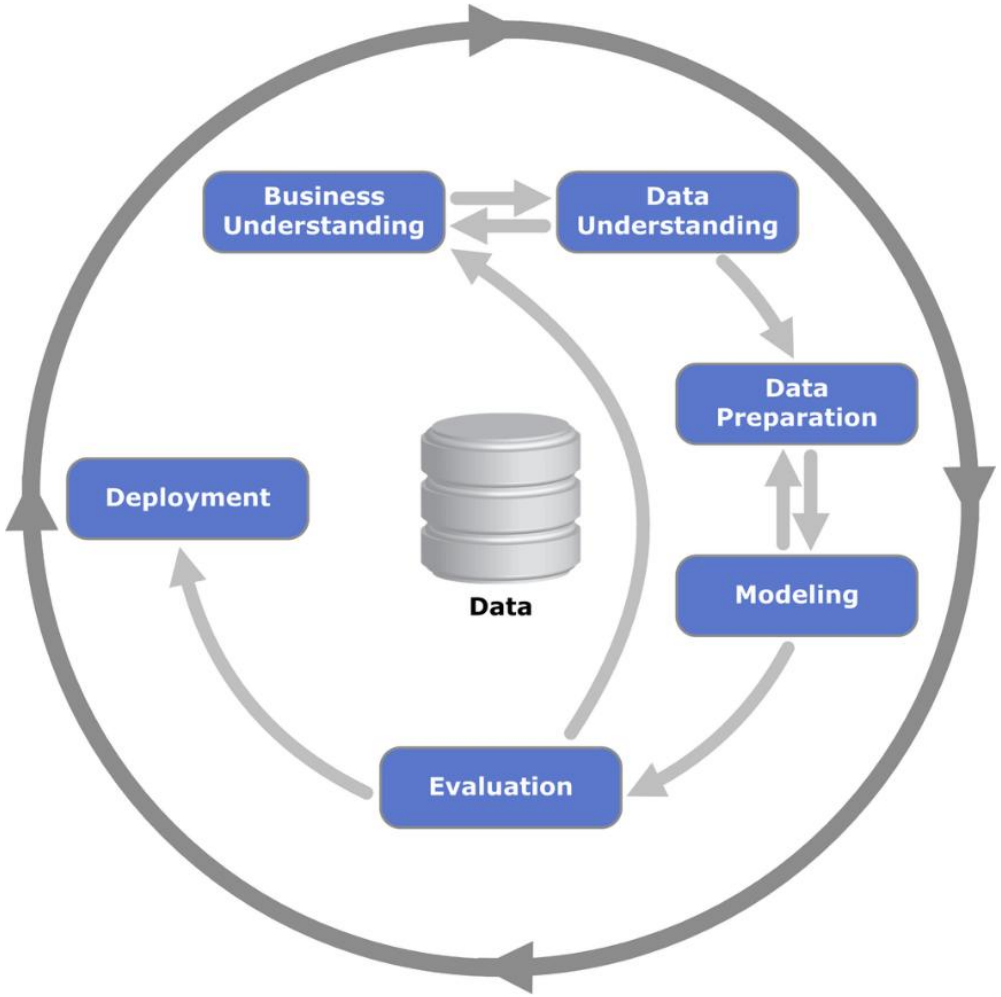
- 평가 데이터에 대한 참고없이 훈련 데이터만으로 최적의 모델을 선택
  - 이론적 선택, 검증 집합, 교차 검증 등의 방식을 사용



데이터 분석 절차

CRISP-DM

- CRISP-DM(Cross-industry standard process for data mining)
  - 데이터 분석을 위한 일반적인 절차에 대한 표준



데이터 분석의 일반적인 절차

- 문제의 설정: 종속변수 Y가 무엇인가?
- 데이터 수집: 관련 데이터 수집 (X, Y)
  - (기존분석) 실험의 설계하고 수행하여 데이터 수집
  - (빅데이터분석) 이미 존재하는 DB에서 관련된 모든 데이터를 수집
- 탐색적 데이터 분석 (EDA, Exploratory data analysis)
  - 데이터에 대해서 배우는 과정
  - 전처리, 시각화, 결측치, 이상치 탐색 등을 포함
- 확정적 데이터 분석 (CDA, Confirmatory data analysis)
  - 클린 데이터로부터 시작 (n = 100M, p = 10k)
  - 훈련 데이터와 평가 데이터를 분리 (보통 시간순서에 따라)
    - 평가 데이터는 훈련이 종료된 후 다시 데이터를 수집하는 것이 가장 좋지만, 실무적으로는 데이터를 한 번 수집하고 둘로 나누어서 사용
  - 학습 모델 후보 선정 (EDA에 따라 4~5개정도 후보 선정)
  - (교차)검증 기법을 이용하여 모델 선정
  - 평가 데이터를 이용하여 최종 성능 평가

- 제3자에 의한 완전히 새로운 데이터로 다시 평가 (필드테스트)

데이터 분석을 배우기 위해서 알아야할 것들

- 탐험적 데이터분석
  - 시각화, 이상치 탐색, 결측치 보정 등
  - 군집분석, 주성분 분석 등
  - 단변량 관계 검정을 통한 종속변수 선택 기법
- 학습모델

	Regression	Classification
Linear Model	Linear Regression	Logistic Regression
Discriminant Analysis		LDA/QDA
Nonparameteric	KNN	KNN, Naïve Bayesian
Tree	Regression Tree	Classification Tree
Ensemble	Bagging, Boosting	
Support Vector	Support Vector Regression	Support Vector Machine
Neural Networks	Multi-layer Perceptron and Deep learning	

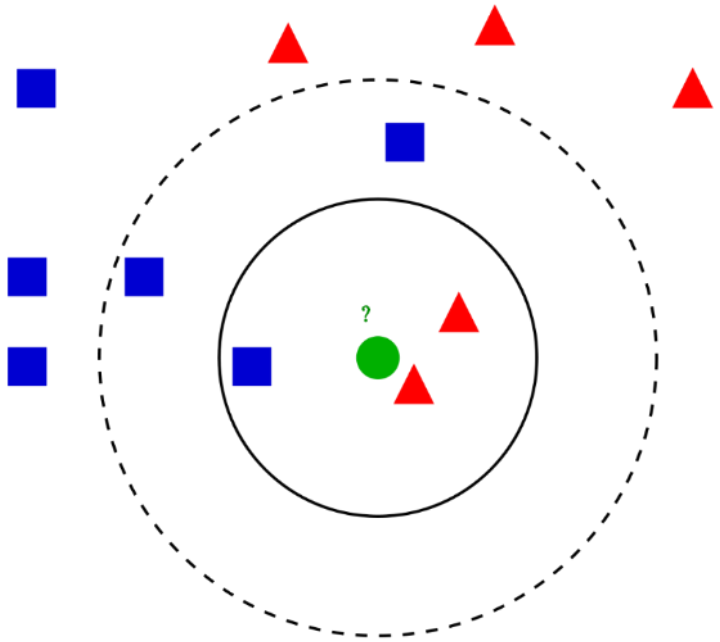
- 모델 선택과 확장
  - 검증 기법, 교차 검증
  - 변수 선택, 벌점화, 차원축소

데이터 분석 절차

최근접 이웃 기법

최근접 이웃 기법 (KNN: K-Nearest Neighbor)

- 예측하려는 지점 주변의 가장 가까운 K개의 샘플로부터 예측
- 회귀와 분류 문제 모두 적용 가능
  - 회귀: 주변 K개 표본의 평균으로 추정
  - 분류: 주변 K개 표본의 최빈값으로 추정 (Major Voting)
- K: 튜닝 파라메터 (하이퍼 파라메터)
  - 사용자가 조절하여 모델의 성격을 변화시킬 수 있는 파라메터



$$\hat{Y} = \frac{1}{K}\sum Y_i$$

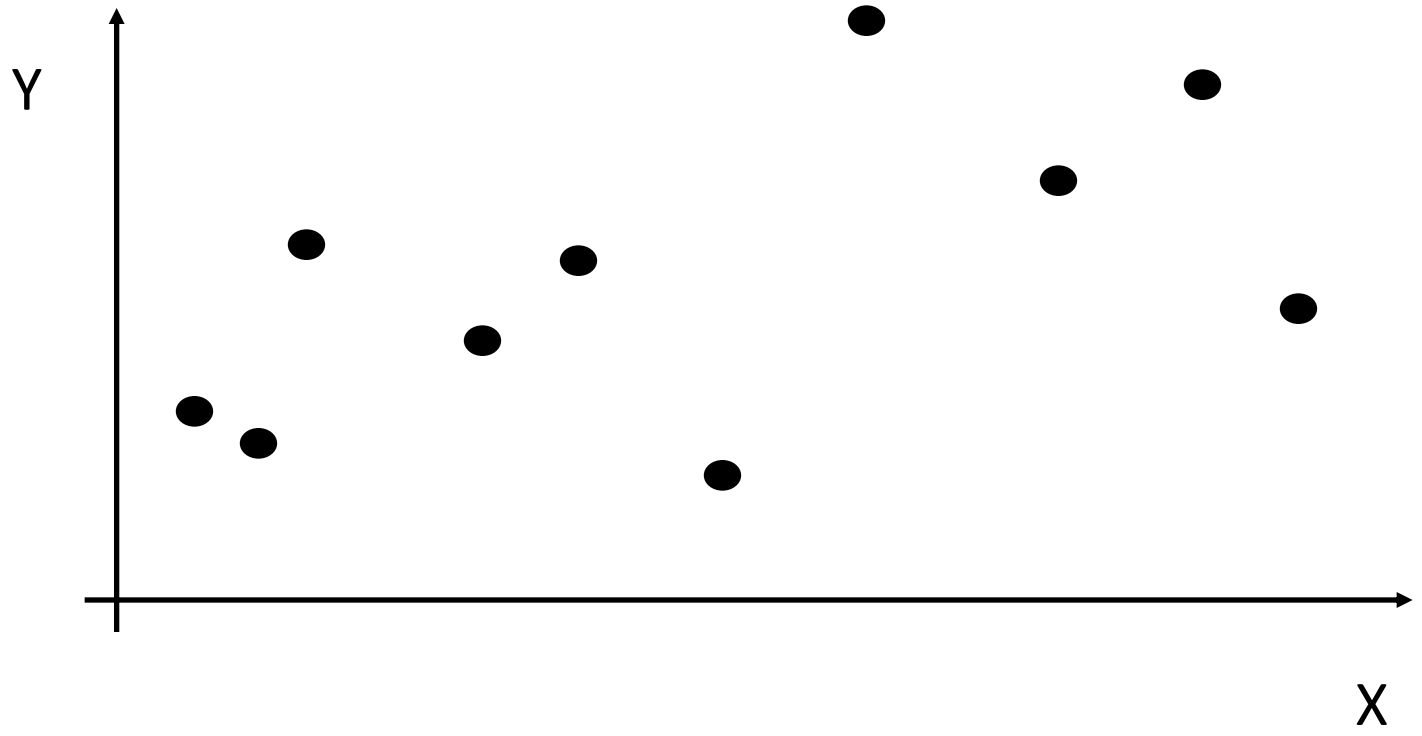
$$\hat{Y} = \underset{c}{\text{argmax}} \sum I(Y_i = c)$$

https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm

최근접 이웃 기법 (KNN: K-Nearest Neighbor)

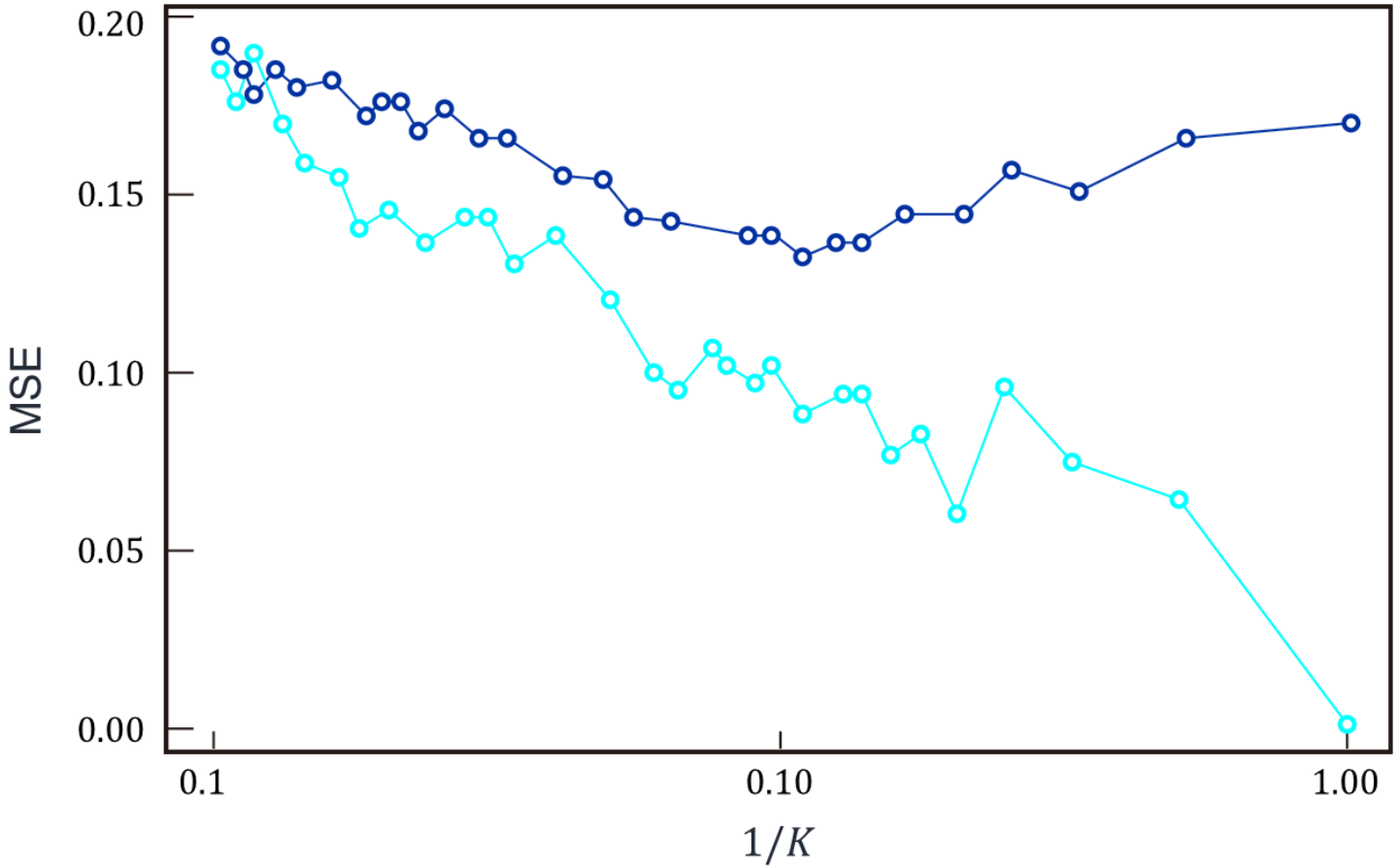
- K에 따른 모델 복잡도의 변화
  - 작은 K: 복잡한 모델
  - 큰 K: 단순한 모델

- K=1 vs. K=10



최근접 이웃 기법 (KNN: K-Nearest Neighbor)

- K에 따른 모델 복잡도의 변화
  - 작은 K: 복잡한 모델
  - 큰 K: 단순한 모델



KOREA UNIVERSITY

감사합니다