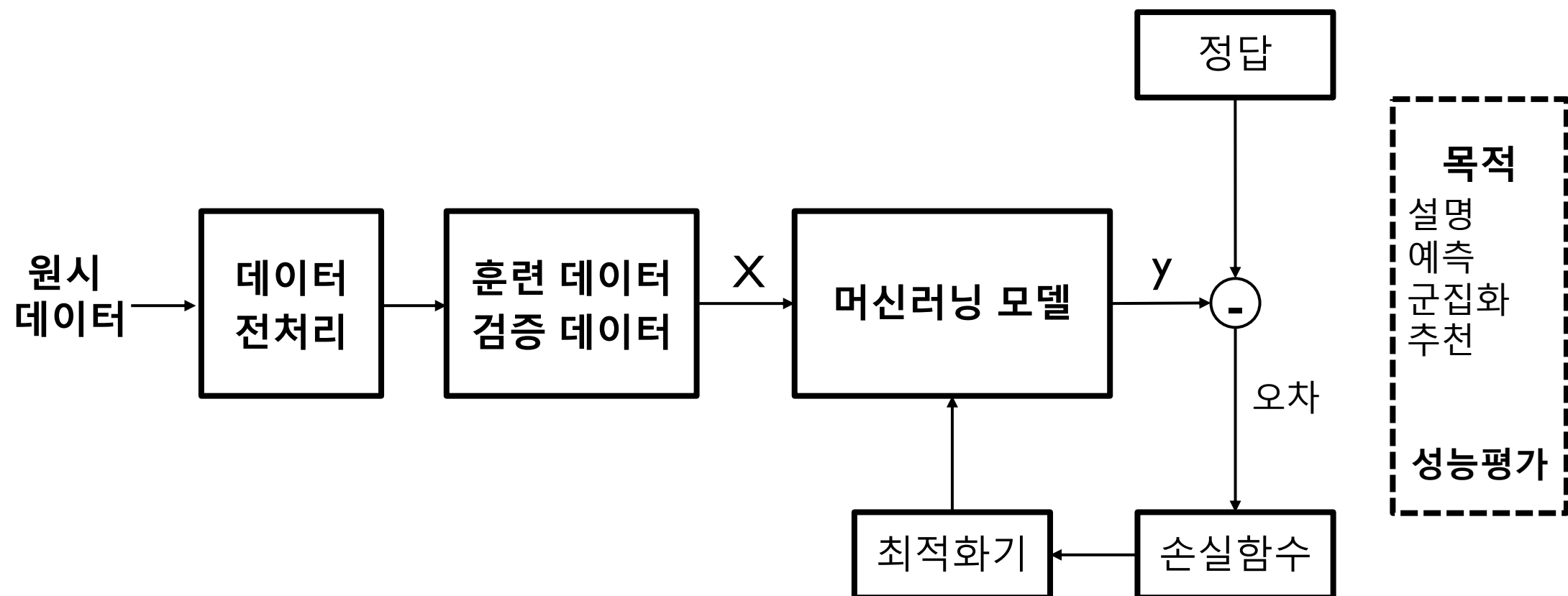


# 머신러닝 이해

# 머신러닝 모델

- ▶ 수식과 논리가 아니라 “데이터 기반”의 모델을 사용한다
- ▶ 훈련/검증 데이터
  - ▶ 모델을 만드는 데는 훈련 데이터를, 성능을 검증하는 데는 검증 데이터를 사용한다

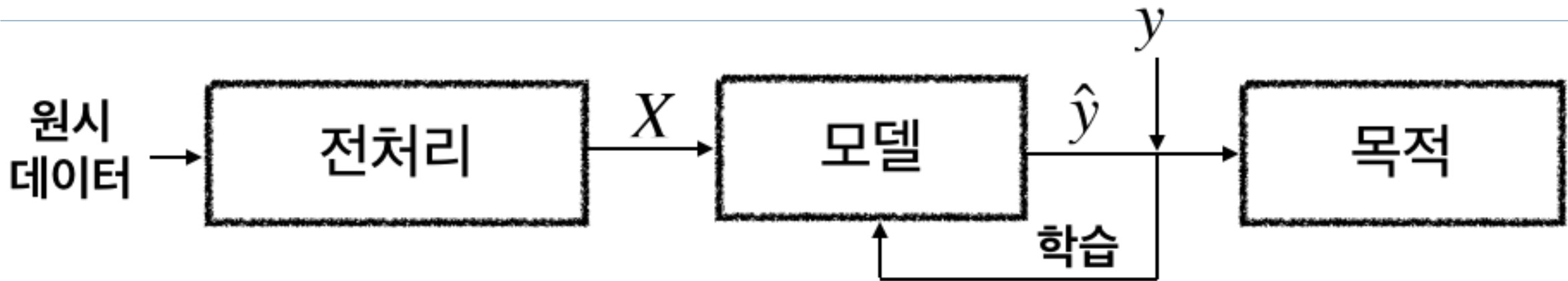


# 손실함수와 성능지표

---

- ▶ **손실함수**를 사용하는 목적은 모델이 얼마나 잘 훈련되는지를 측정하기 위해서이다.
  - ▶ 모델은 손실함수를 최소화 하는 방향으로 학습한다.
- ▶ **모델의 성능지표**는 모델이 궁극적으로 얼마나 잘 동작하는지를 평가하는 척도이다.
  - ▶ 성능이 높은 모델을 만드는 것이 목적이다

# 머신러닝 프로세스



결측치 처리  
오류값 처리  
스케일링  
데이터 변환  
· 카테고리 변환  
· 로그, 역수 변환  
특성공학  
· 차원축소  
· PCA

선형모델  
로지스틱회귀  
**SVM**  
결정트리  
랜덤포레스트  
그라디언트부스트  
**kNN, Bayes**  
**CNN**  
**RNN**

클러스터링  
설명적 분석  
· EDA, 시각화  
회귀 예측  
분류 예측  
추천

# 머신러닝 모델 특징

머신러닝 유형		알고리즘	특징
지도학습 (예측모델)	선형 계열	선형 모델, SVM 로지스틱회귀	곱셈과 덧셈으로 점수를 구하고 이를 이용하여 회귀와 분류 예측
	신경망	MLP, CNN, RNN, Transformer	매트릭스 연산을 기반으로 점수를 계산하며 활성화 함수 도입
	트리 계열	결정 트리, 랜덤포레스트, 그라디언트부스팅	True/False 선택을 반복하여 회귀와 분류 예측 수행. 스케일링이 필요없다
	기타	kNN, 베이지	특성 공간상의 거리를 기준, 또는 조건부 확률을 기준으로 예측
비지도학습 (데이터 처리)	클러스터링	k-means, DBSCAN	특성 공간상 거리와 유사도를 기준으로 샘플을 그룹핑
	데이터 변환	스케일링, 로그변환, 카테고리 인코딩	효과적인 데이터 전처리
	차원 축소	PCA, t-SNE	계산량과 모델 성능 향상, 의미 있는 시각화

# 머신러닝 모델 유형

---

- ▶ 선형계열 모델
  - ▶ 입력 특성(features)들에 대해 가중합(weighted sum) 연산을 수행하여 점수를 구하고 이 점수를 사용하여 회귀 및 분류를 수행한다
- ▶ 트리계열 모델
  - ▶ 연산을 수행하지 않고, 특성별로 조건식을 적용하여 이진 분류를 순차적으로 수행하여 회귀 및 분류를 수행한다
- ▶ 신경망 모델
  - ▶ 선형계열 모델처럼 가중합 연산을 수행한 점수를 사용하되, 매우 크고 다양한 가중 매트릭스를 사용하여 성능을 개선한다

# 손실함수와 성능지표

## ▶ 대표적인 손실함수와 성능 평가 지표

	손실함수	성능평가지표
정의	손실함수를 줄이는 방향으로 모델이 학습을 함	성능을 높이는 것이 머신러닝을 사용하는 최종 목적임
회귀 모델의 대표적인 값	MSE (Mean Squared Error)	$R^2$
분류 모델의 대표적인 값	크로스 엔트로피	정확도, 정밀도, 재현률, F1점수, ROC-AUC

# 과대적합

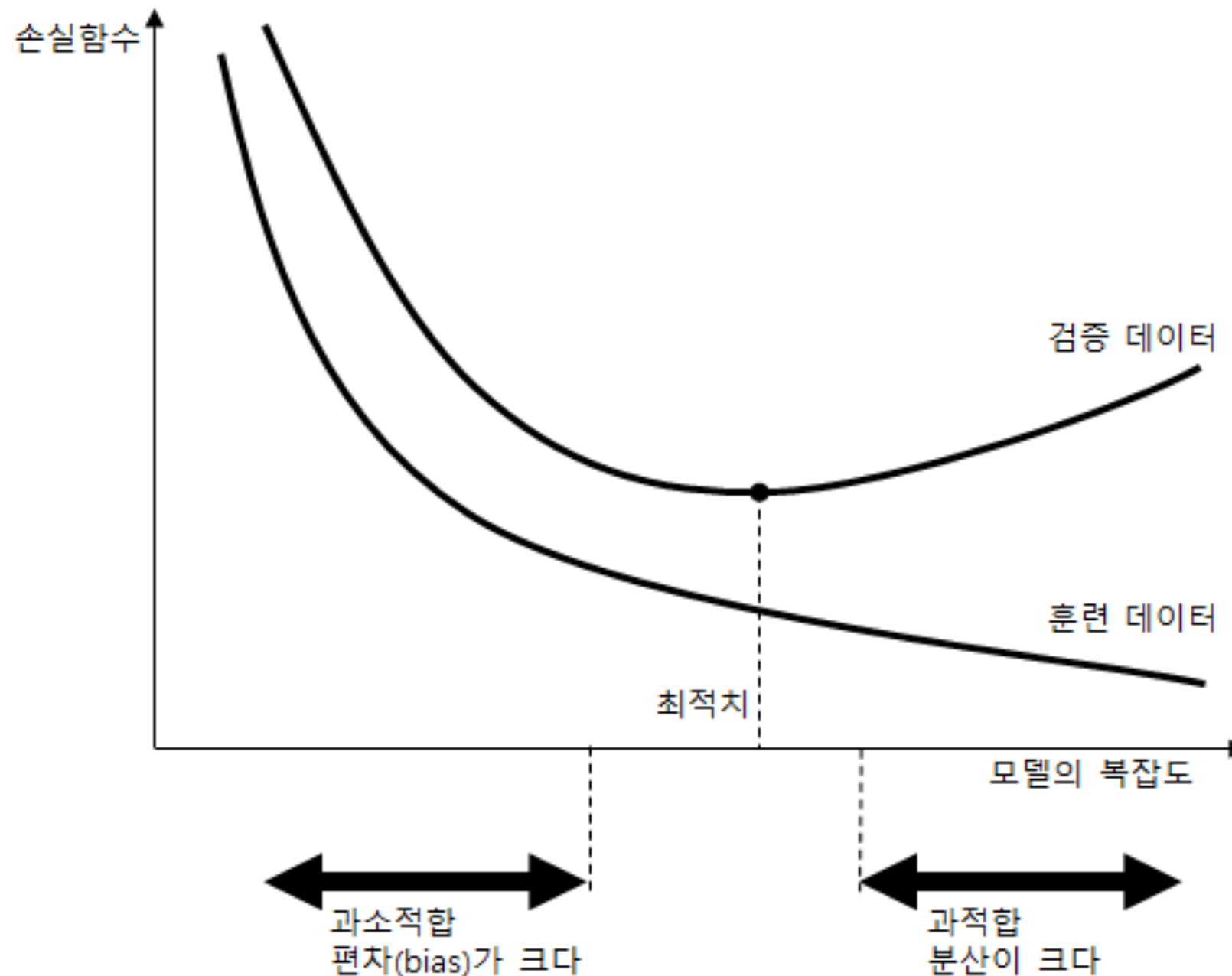
---

- ▶ 모델이 훈련 데이터에 대해서는 잘 동작하지만 새로운 데이터에 대해서는 오히려 잘 동작하지 못하는 경우를 과대적합(over fitting)되었다고 한다.
- ▶ 과대적합은 주어진 훈련 데이터를 너무 세밀하게 학습에 반영하여 발생하는 현상이다.
- ▶ 과대적합을 줄이려면 더 많고 다양한 학습 데이터를 사용하거나 모델에 제한을 두어 일반화하는 것이 필요하다



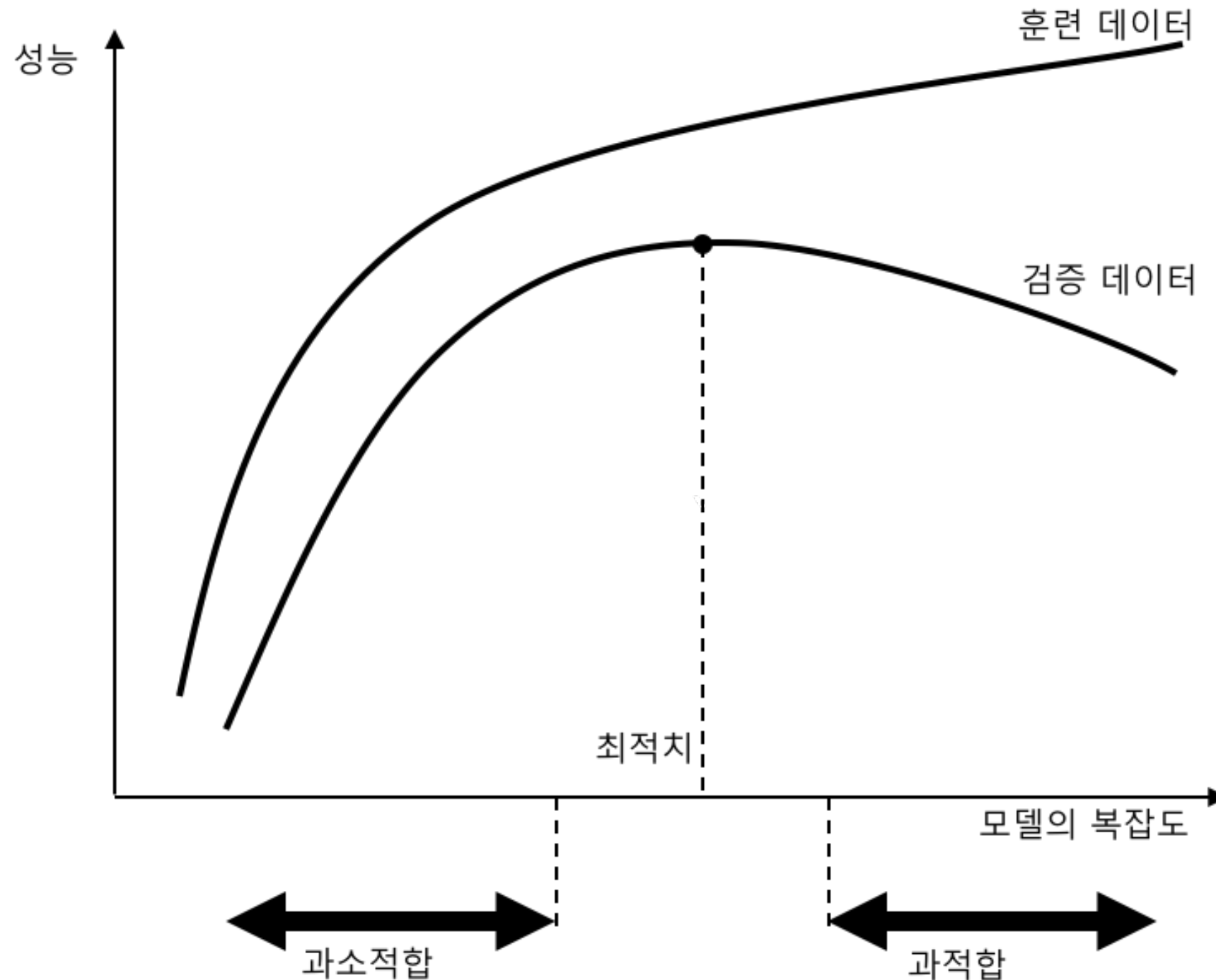
# 과대적합 검증

- ▶ 훈련 데이터와 검증 데이터에 대한 손실함수를 비교한다

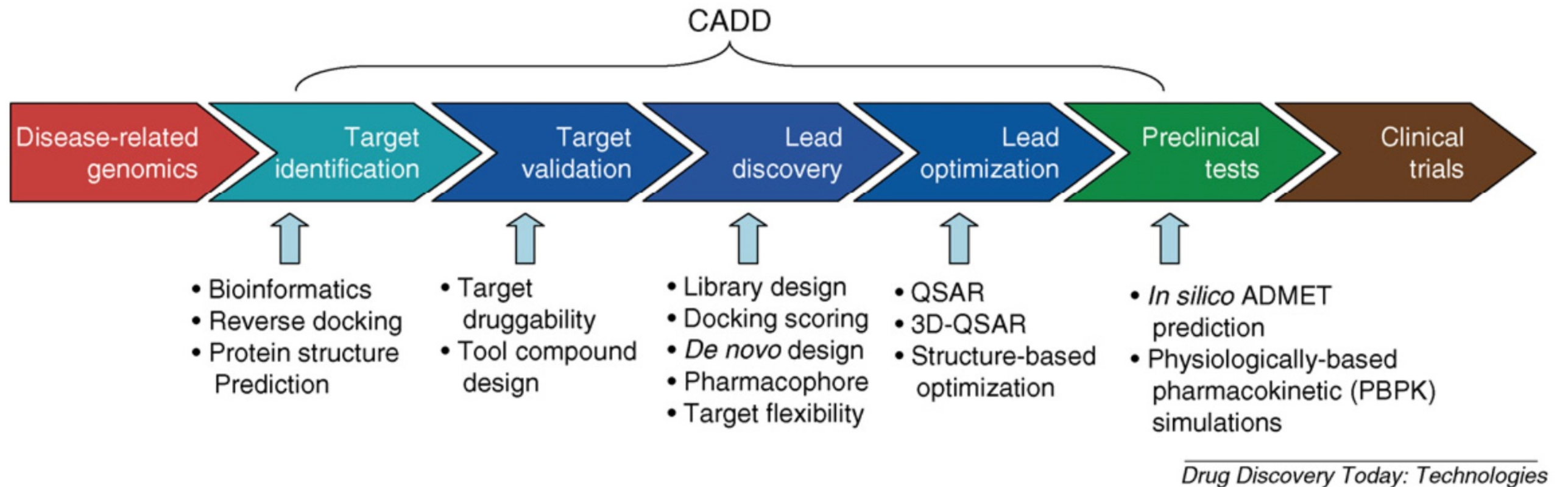


# 과대적합 검증

- ▶ 훈련 데이터와 검증 데이터에 대한 성능을 비교한다



# CADD with ML



## ▶ Machine Learning Models

### ▶ Traditional models (1990~)

- ▶ Linear, logistic regression, support vector machines (SVM)
- ▶ Decision Tree, Random Forest, Boosting

### ▶ Deep Neural Networks (2012~)

- ▶ MLP, CNN, RNN, Graph NN, Transformer