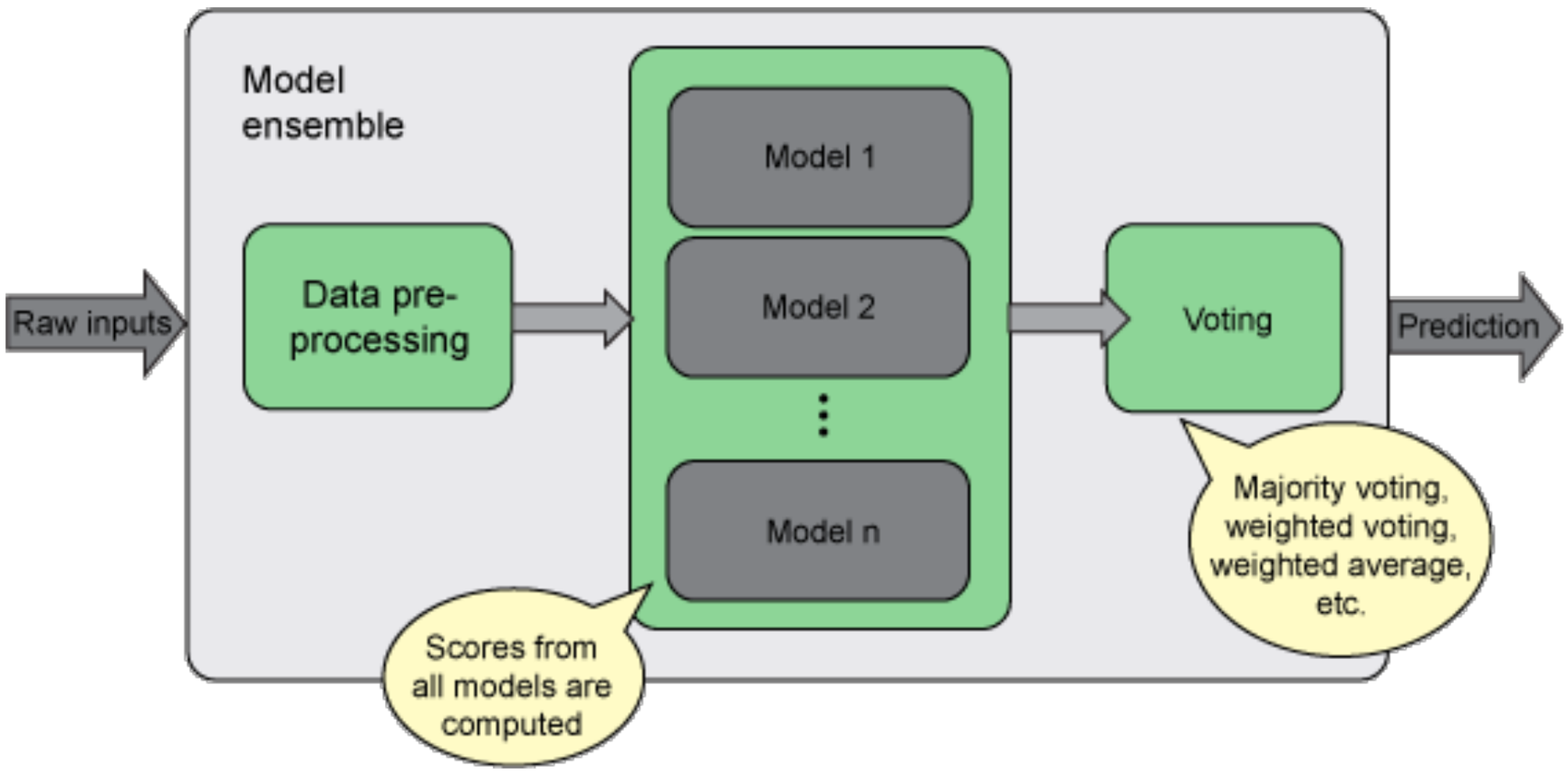
Ensemble

Ensemble

여러 모델을 이용하여 결과를 예측하고, 이를 결합하는 것을 Ensemble이라고 합니다



- **Error 최소화**

다양한 모델의 결과를 종합하여 전반적 오류를 줄여줌

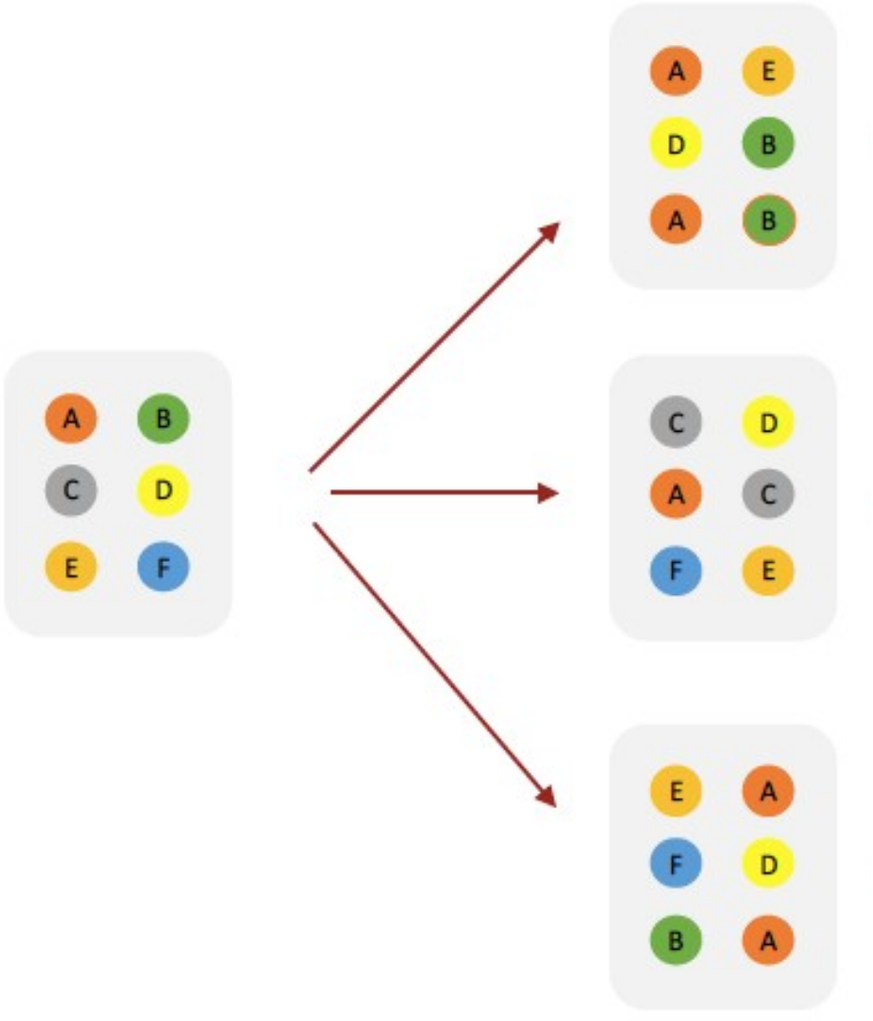
- **모델별 한계를 극복**

모델별로 가지고 있는 한계를 여러가지 모델의 결과를 종합해 극복

Bagging & RandomForest

Bagging

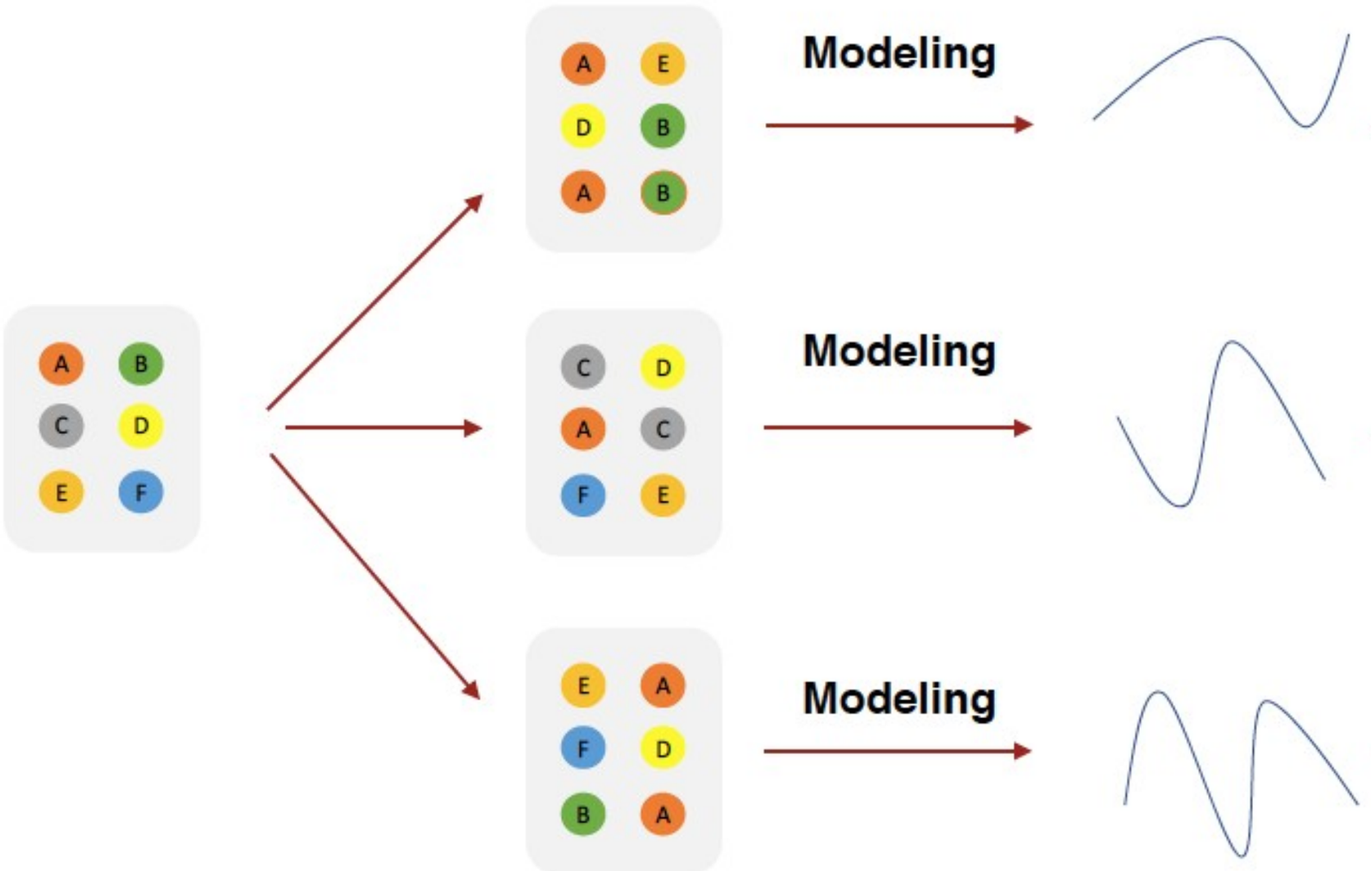
Bagging은 Bootstrap Aggregating의 줄임말입니다



Random Sampling with replacement

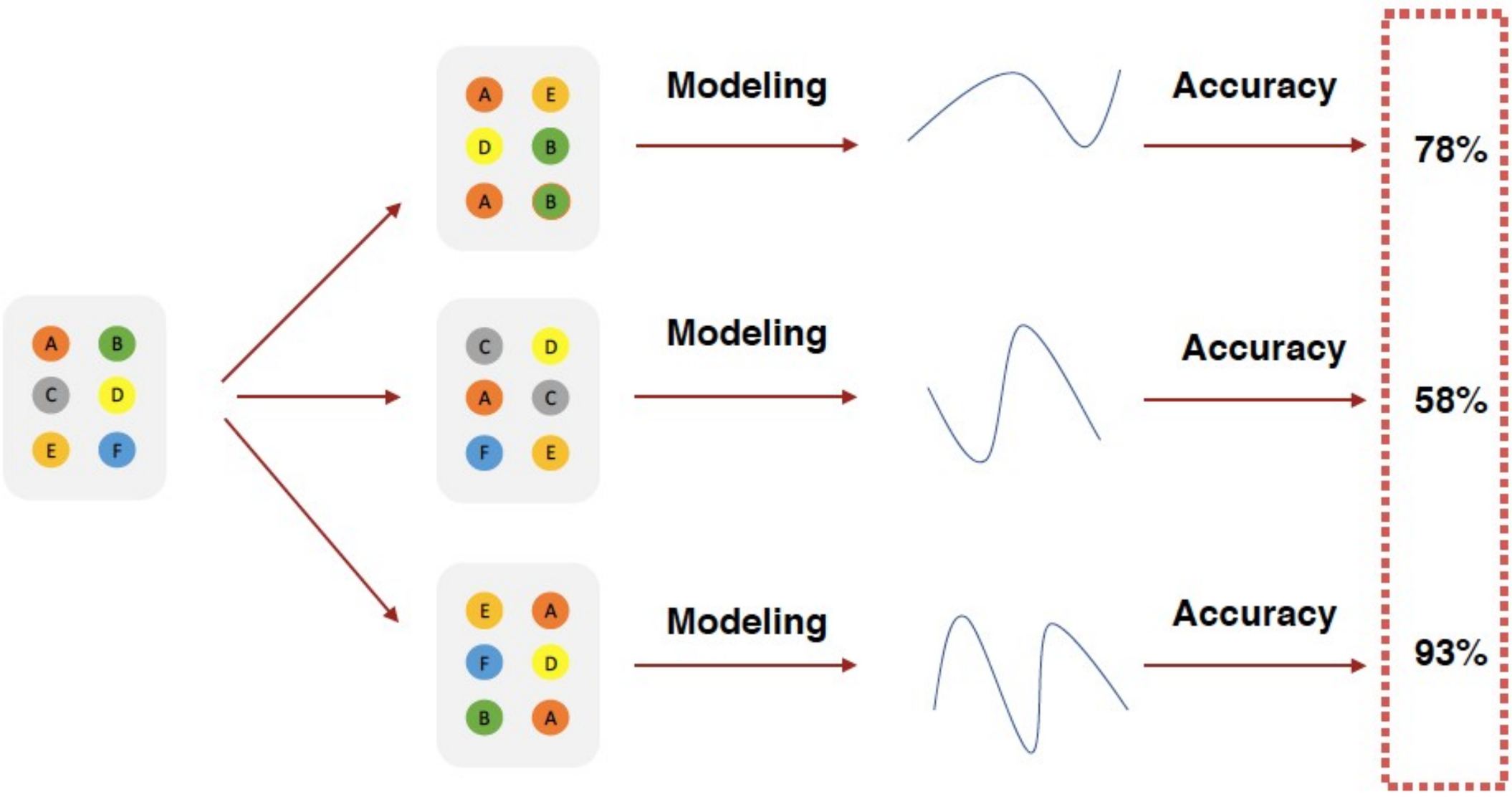
Bagging

추출된 표본들에 각각 모델 (ex. Decision Tree)를 적합시켜 모델을 만들 수 있습니다



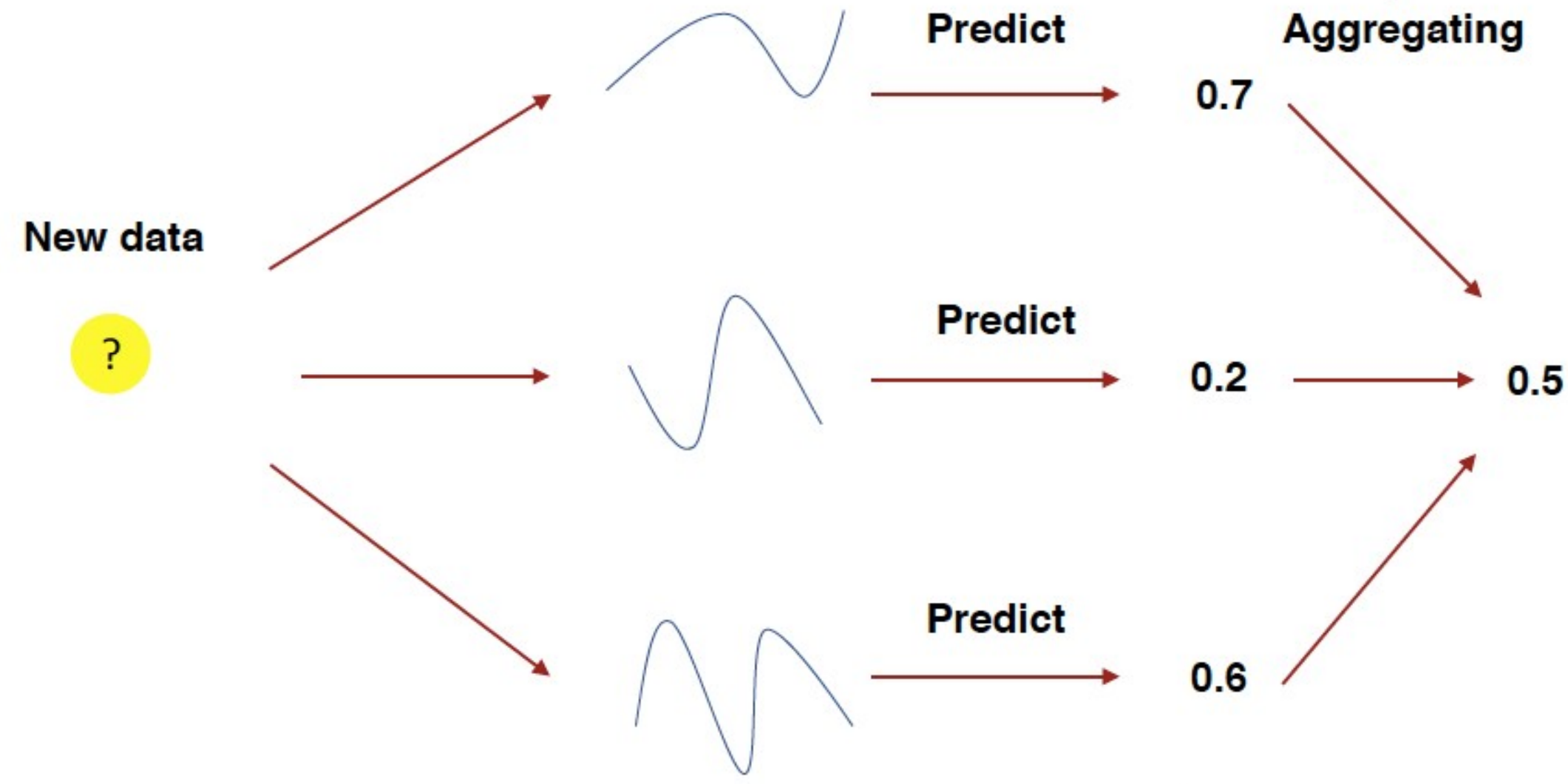
Bagging

그리고 각각의 모델들은 예측 성능을 가지고 있습니다
이 값을 바탕으로 예측 성능을 얼마나 믿을 수 있는지 측정할 수 있습니다



Bagging

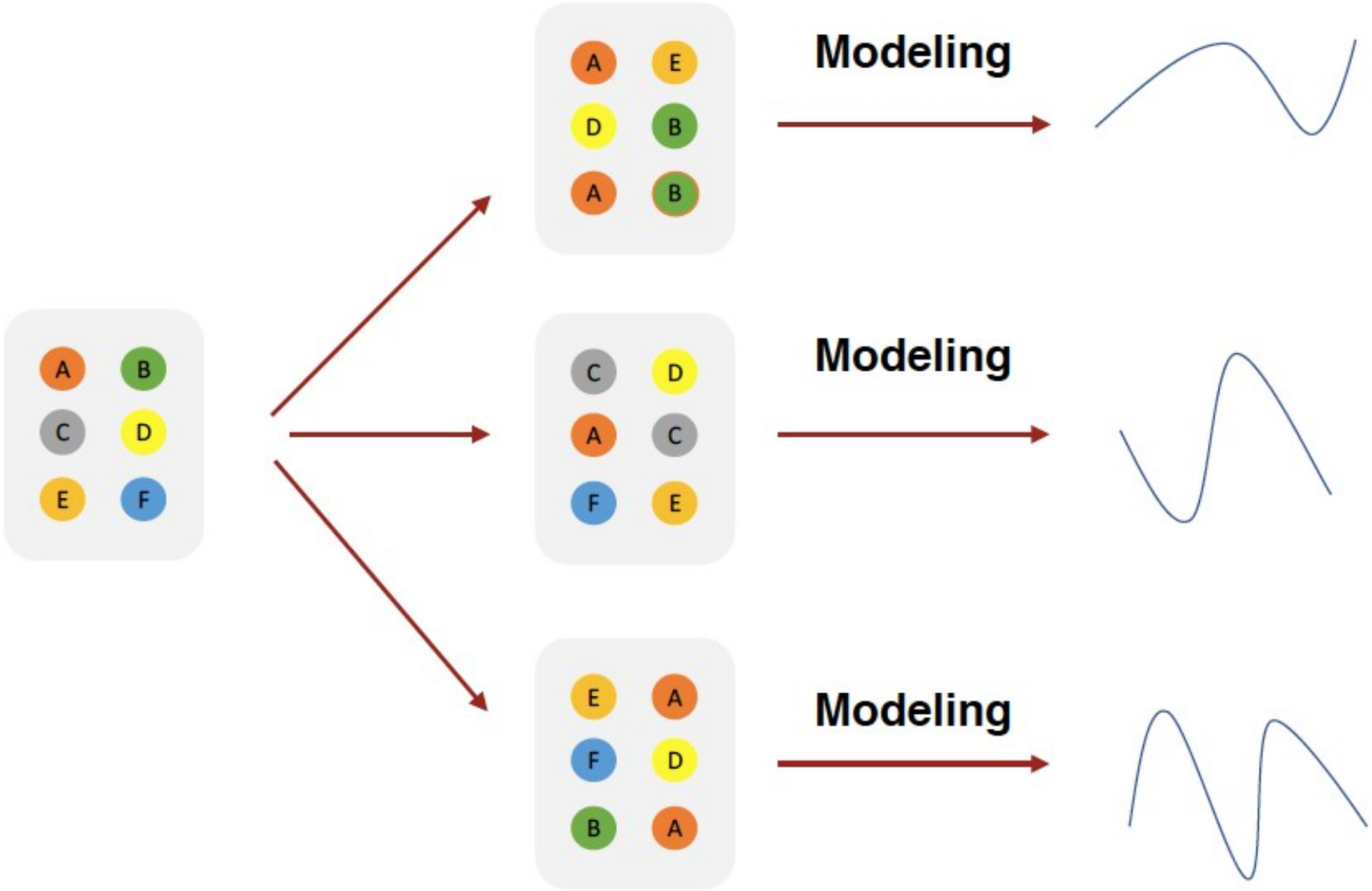
각각의 예측 결과를 합쳐 하나의 예측결과로 만들 수 있습니다



Classification : 다수결
Regression : 평균

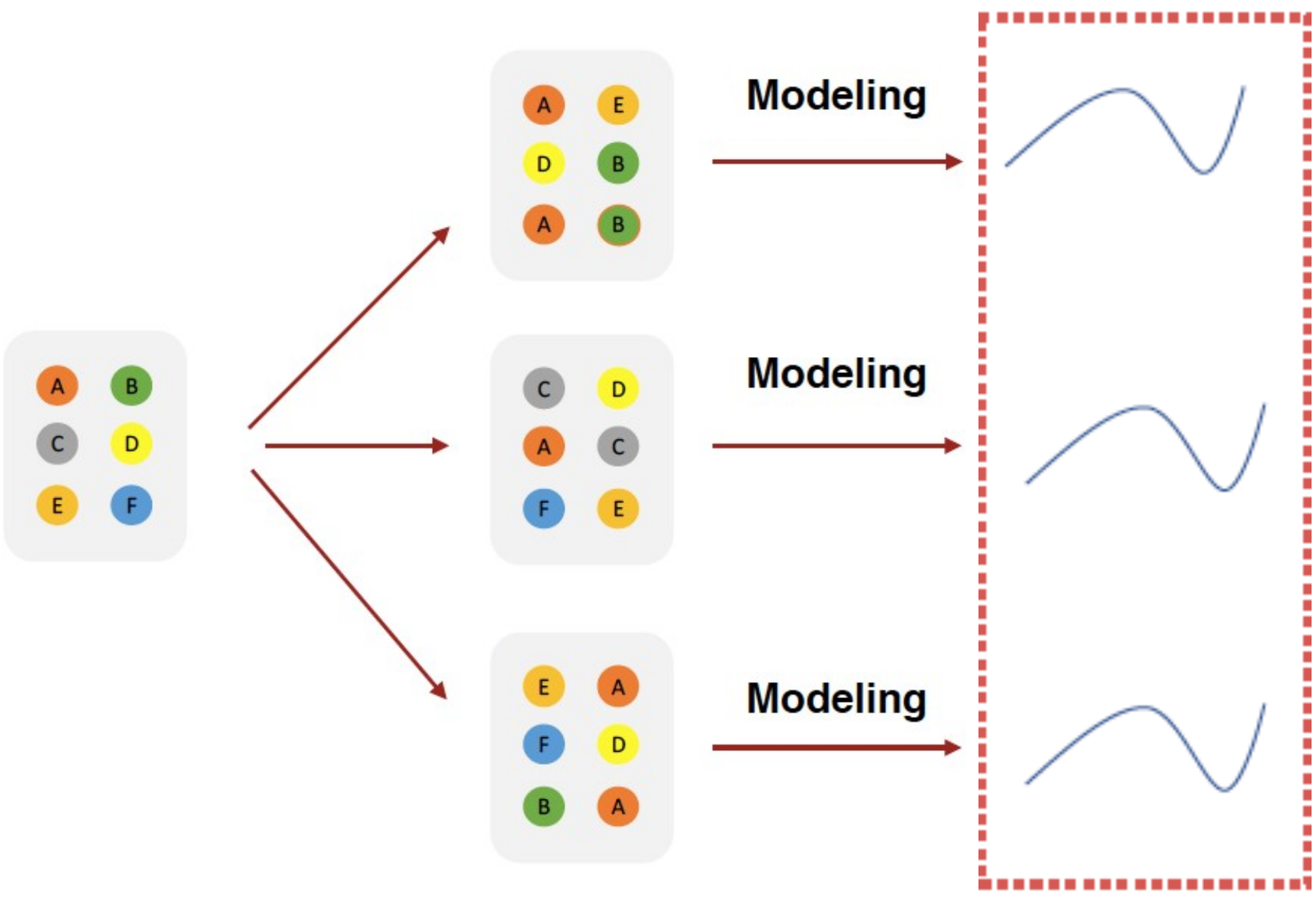
Bagging의 문제점

다음과 같이 같은 데이터에 대해 다양한 모델이 나오는 것이 이상적입니다



Bagging의 문제점

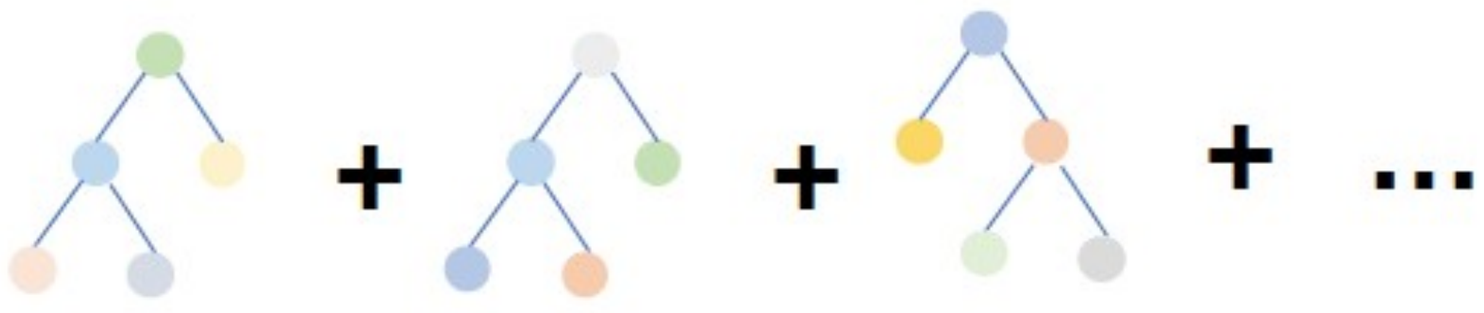
하지만 다음과 같이 모델이 전부 비슷하게 만들어지면 합치는 의미가 없어집니다



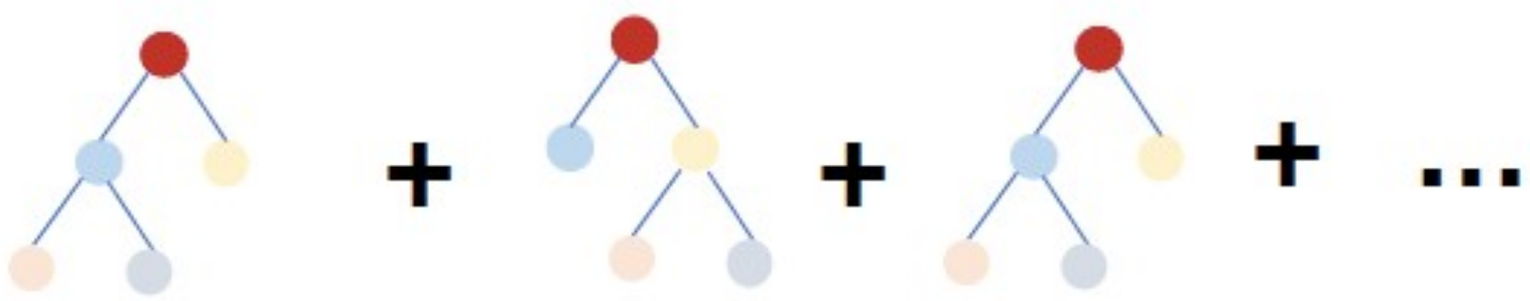
Bagging의 문제점

강력한 설명변수가 있다면 tree가 그 변수로 인해 대부분 비슷해집니다

이상적인 Bagging



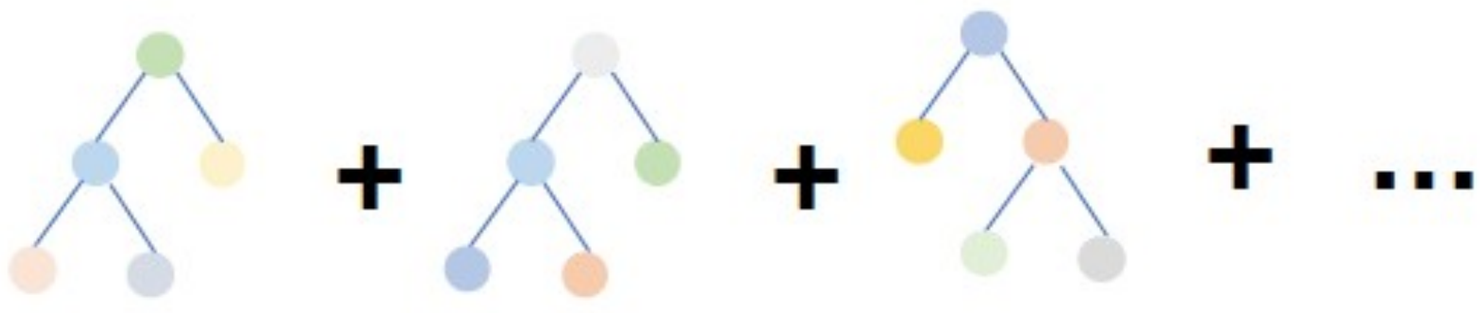
강력한 설명변수가 있을 때



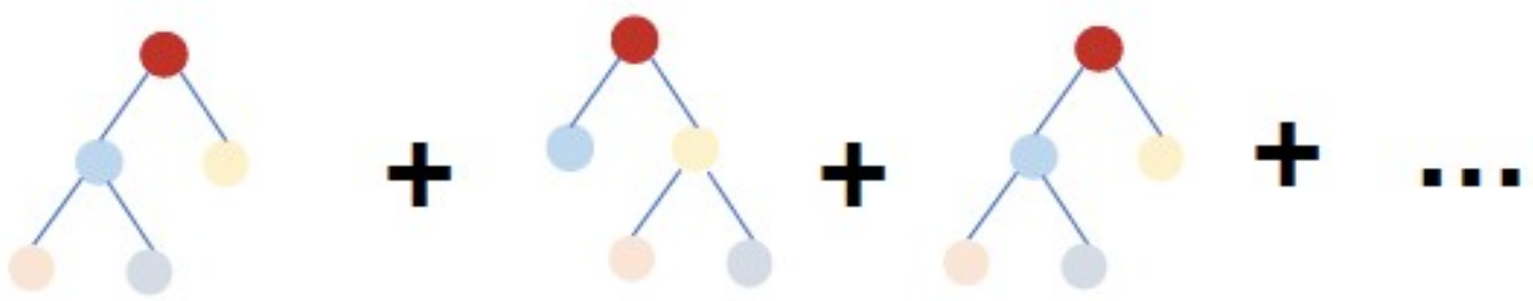
Bagging의 문제점

강력한 설명변수가 있다면 tree가 그 변수로 인해 대부분 비슷해집니다

이상적인 Bagging



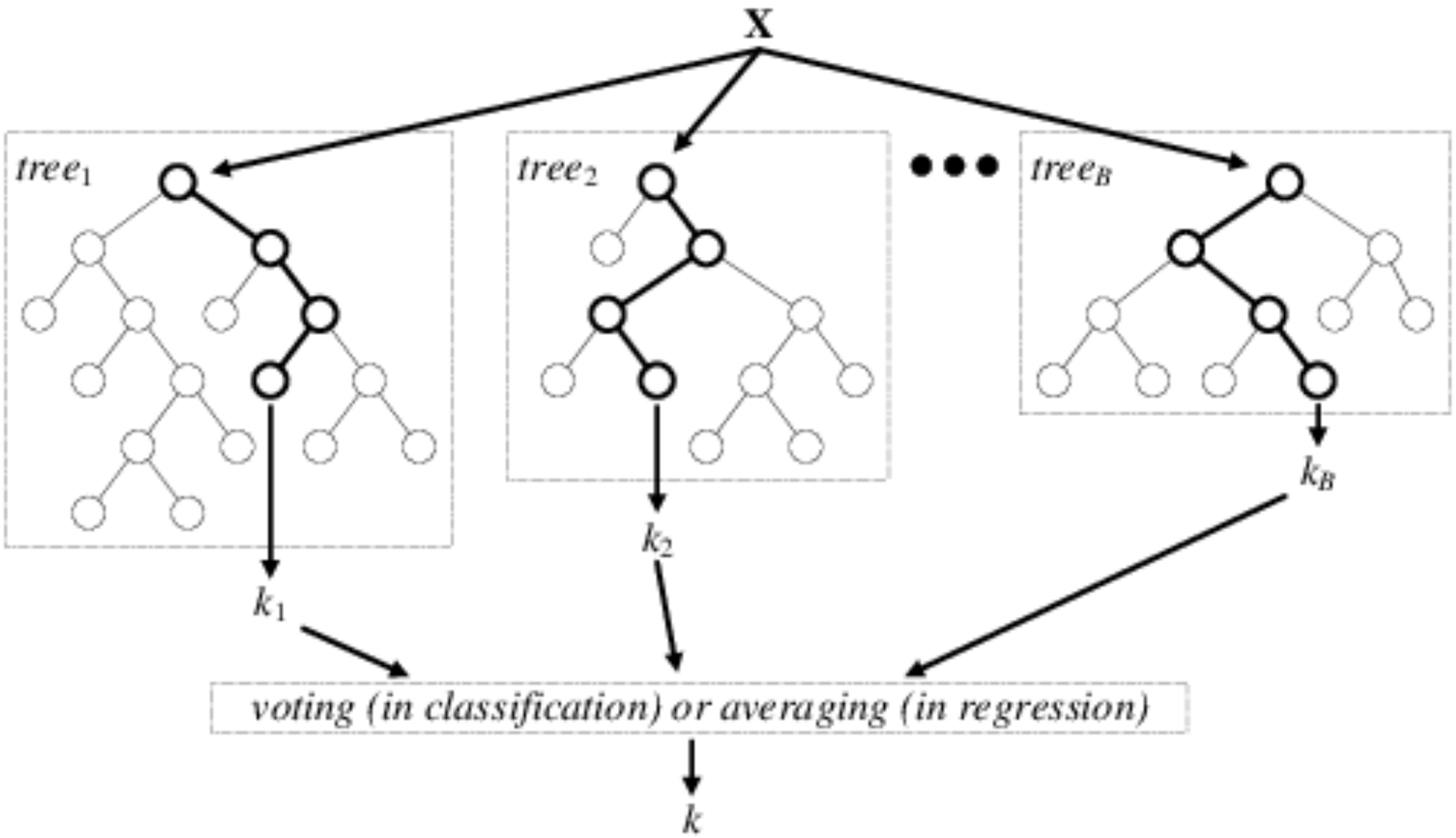
강력한 설명변수가 있을 때



트리간에 강력한 상관관계가 생겨 이를 ensemble해도 분산이 감소하지 않습니다

Random Forest

RandomForest는 Bagging과 동시에 각 tree마다 사용할 수 있는 최대 변수의 개수를 제한합니다

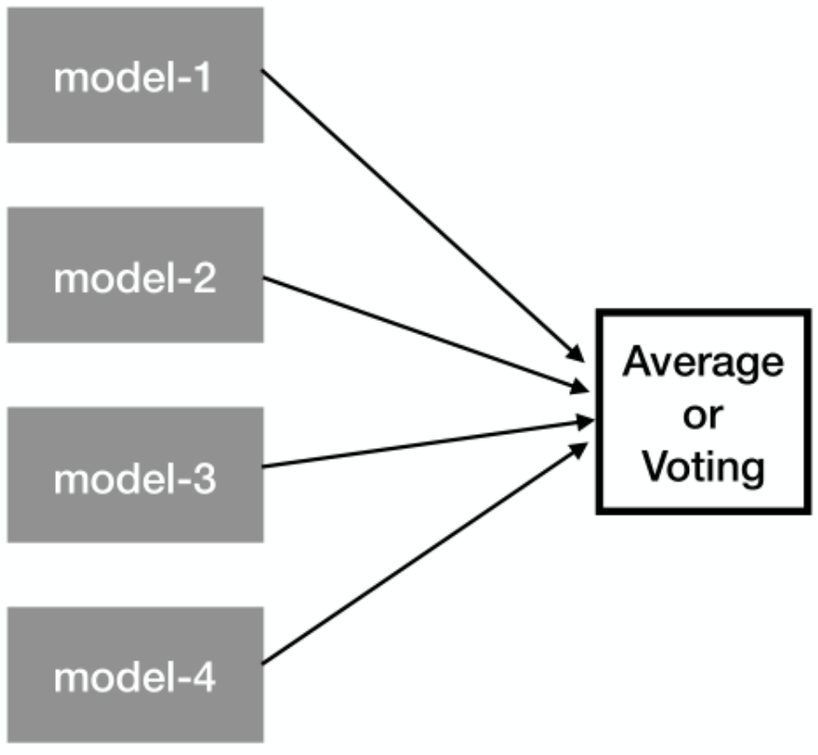


최대 변수의 수는 일반적으로 $\sqrt{\text{총 변수의 수}}$를 사용합니다

Boosting

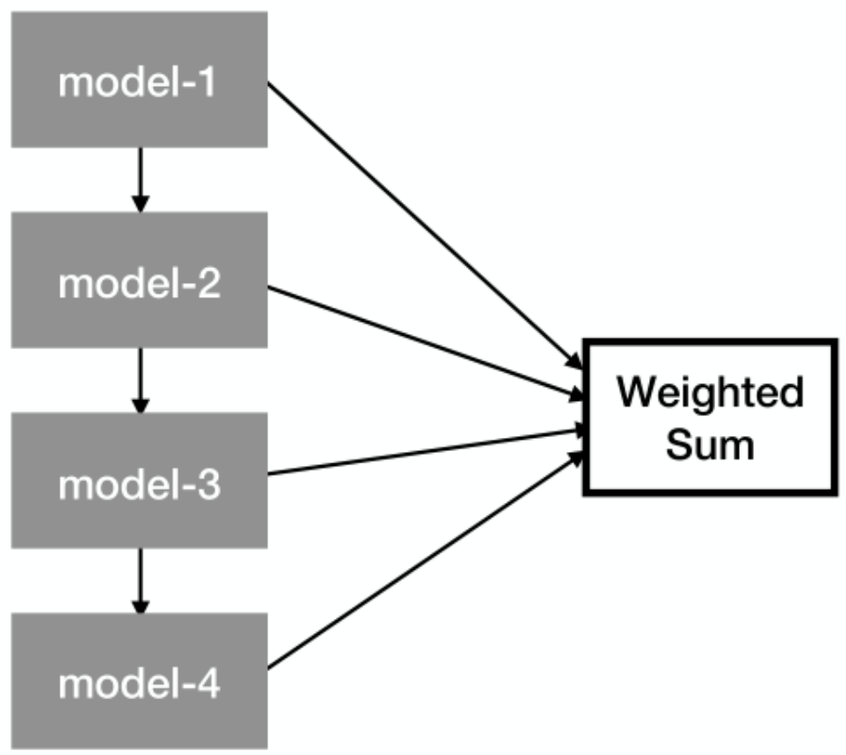
Bagging vs Boosting

Bagging



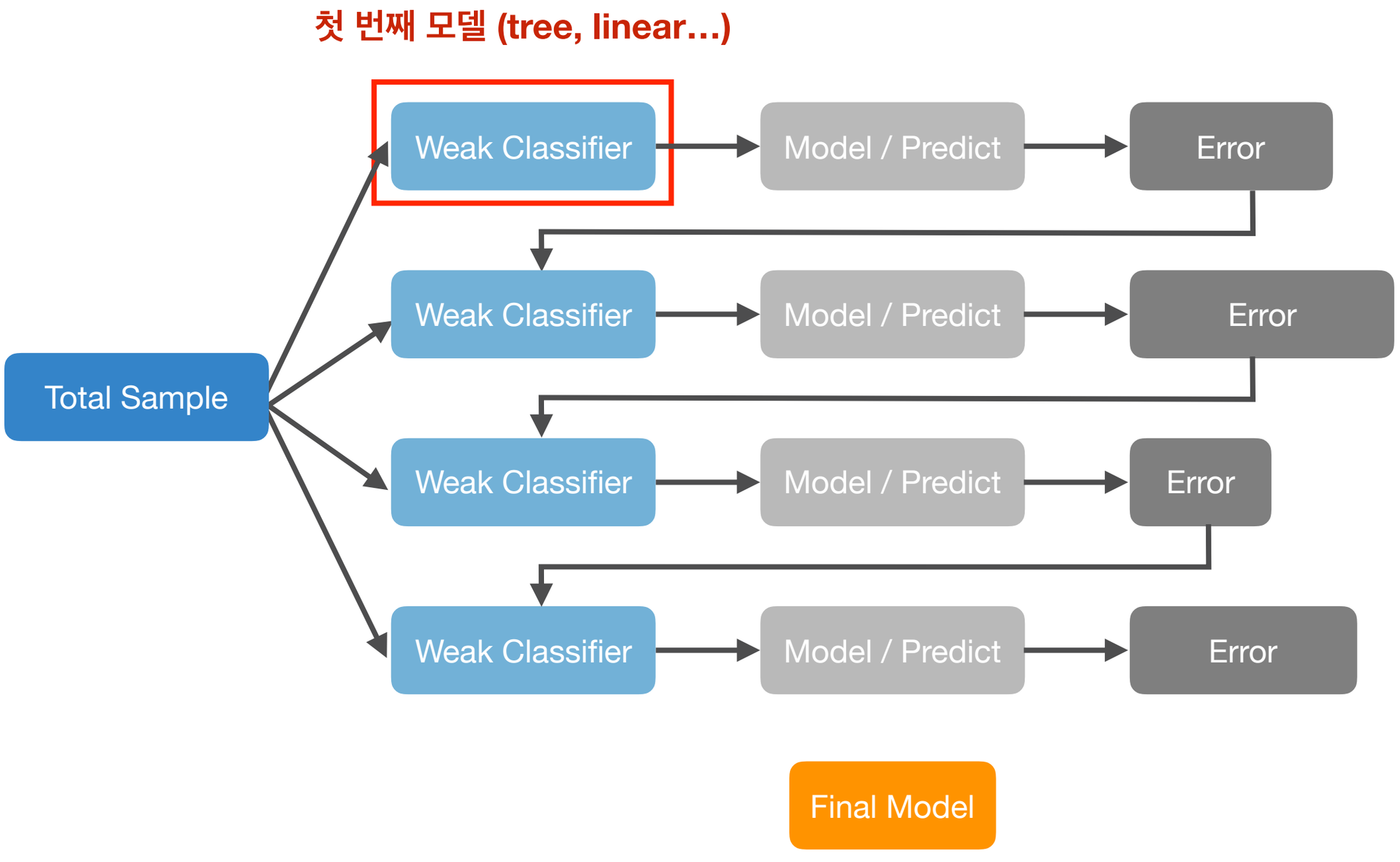
각 모델들은 병렬적으로 학습
모델끼리 영향을 주지 않음

Boosting

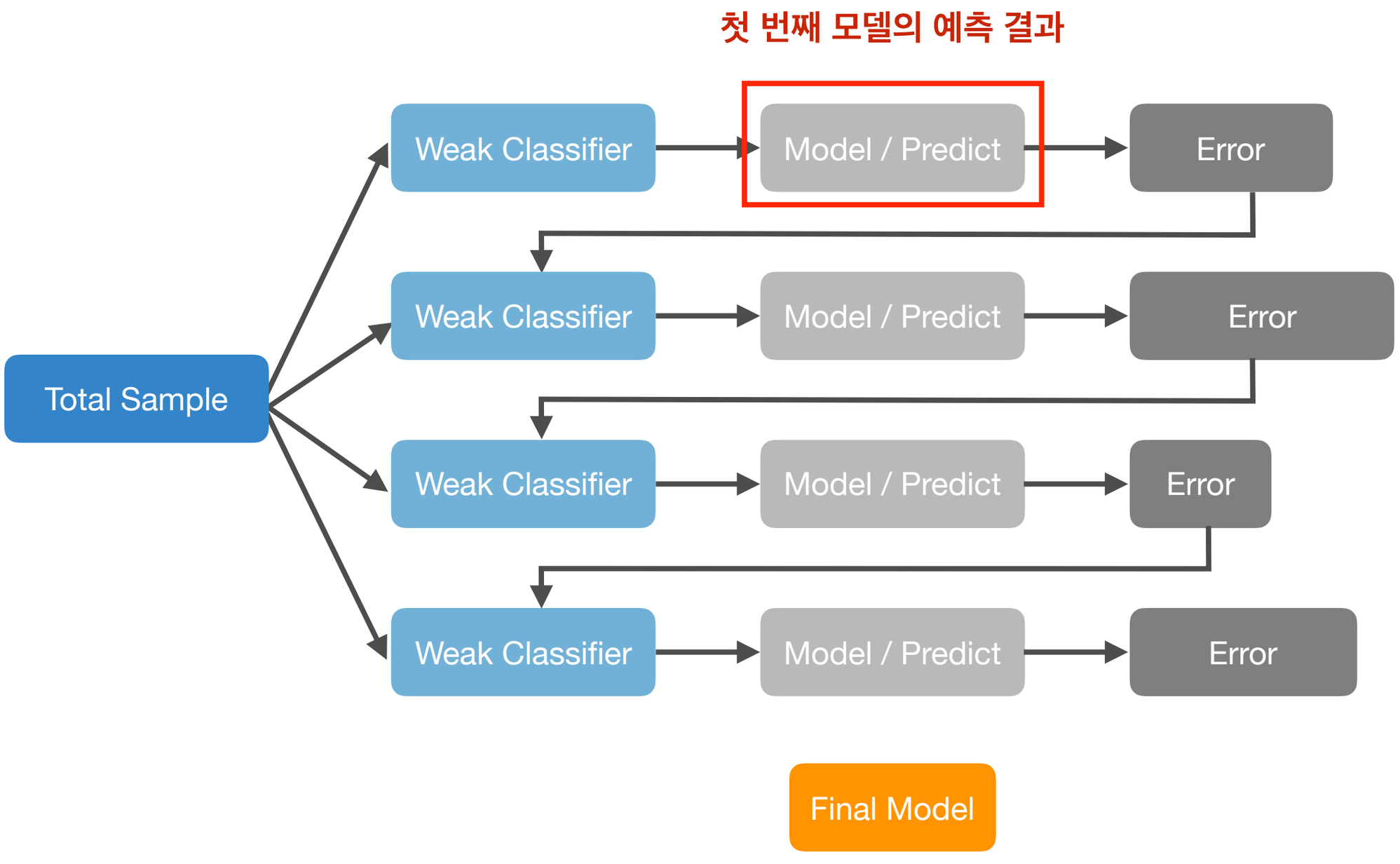


각 모델들은 순차적으로 학습
이전 모델의 학습 결과를 바탕으로 다음 모델을 학습

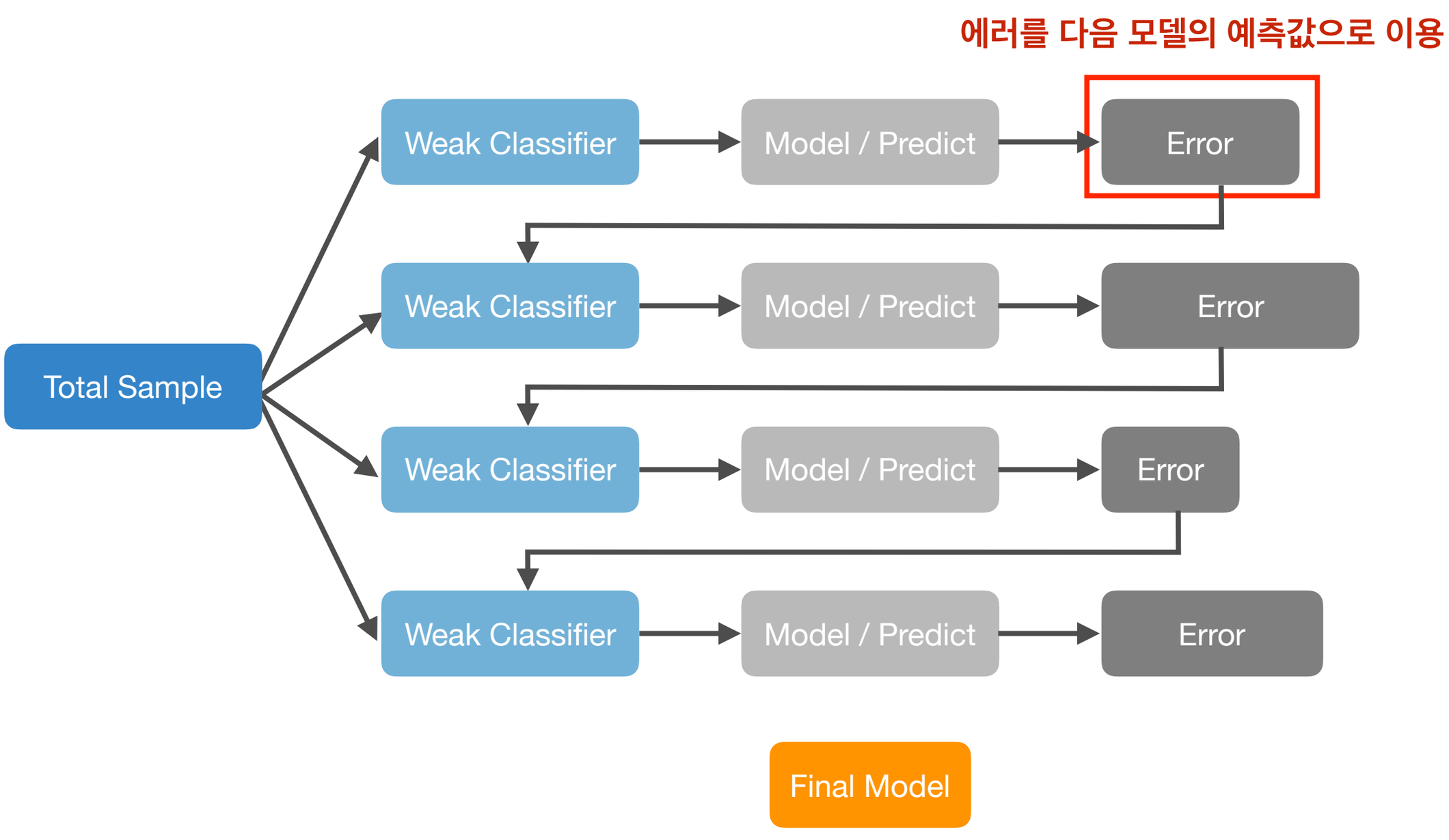
Gradient Boosting 작동 원리



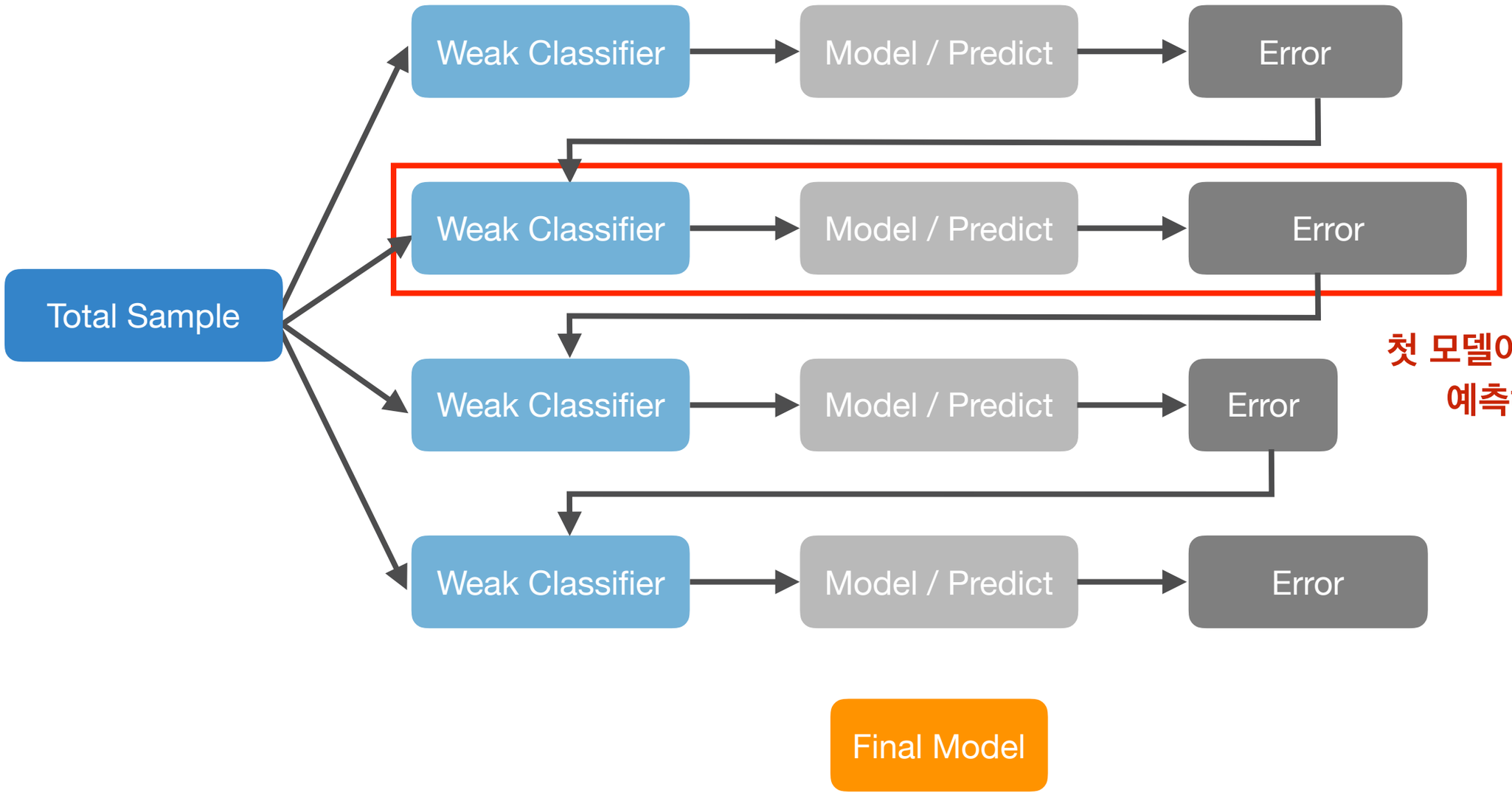
Gradient Boosting 작동 원리



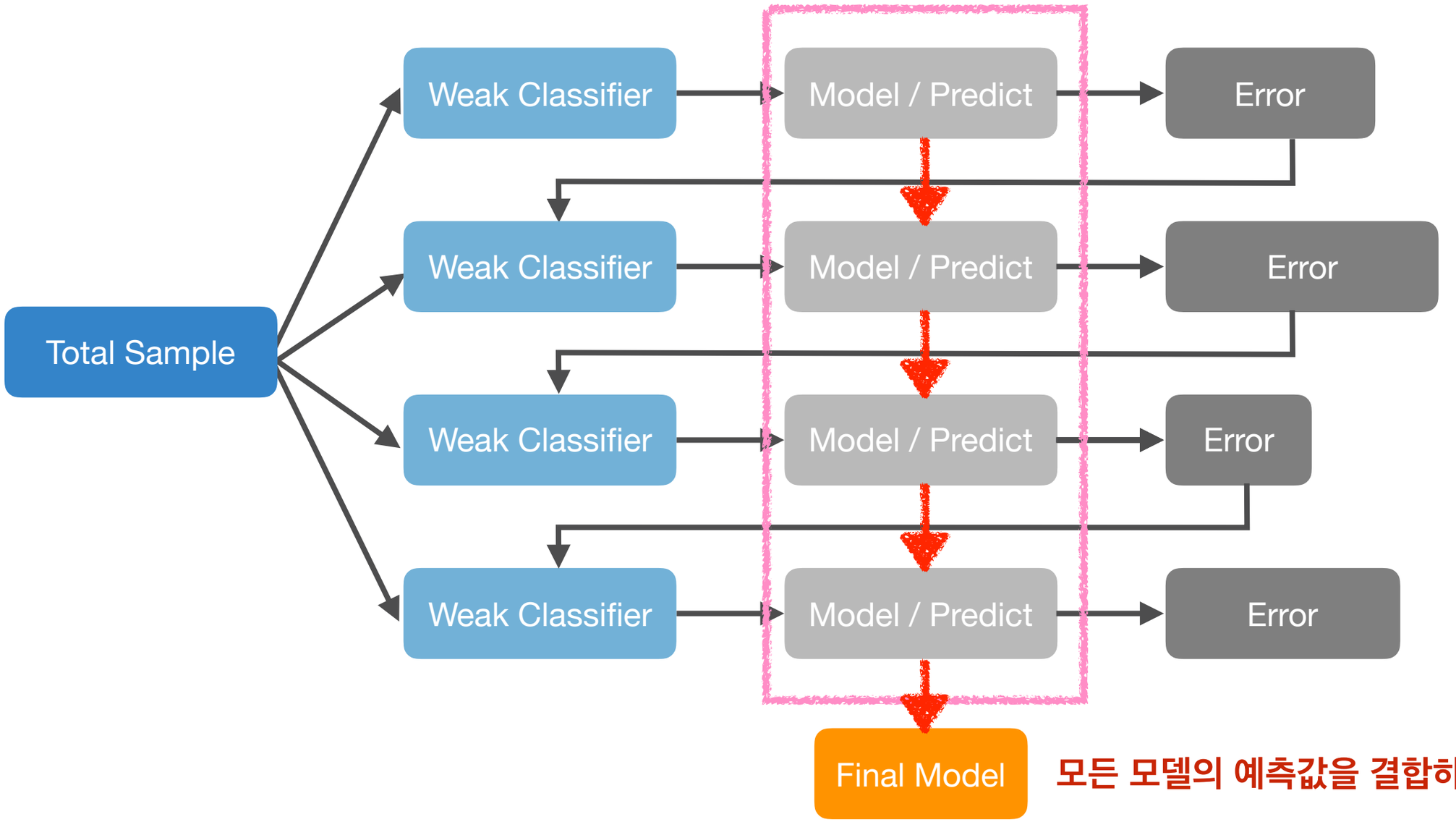
Gradient Boosting 작동 원리



Gradient Boosting 작동 원리



Gradient Boosting 작동 원리



모든 모델의 예측값을 결합하여 최종 모델 생성