모델 평가

모델?

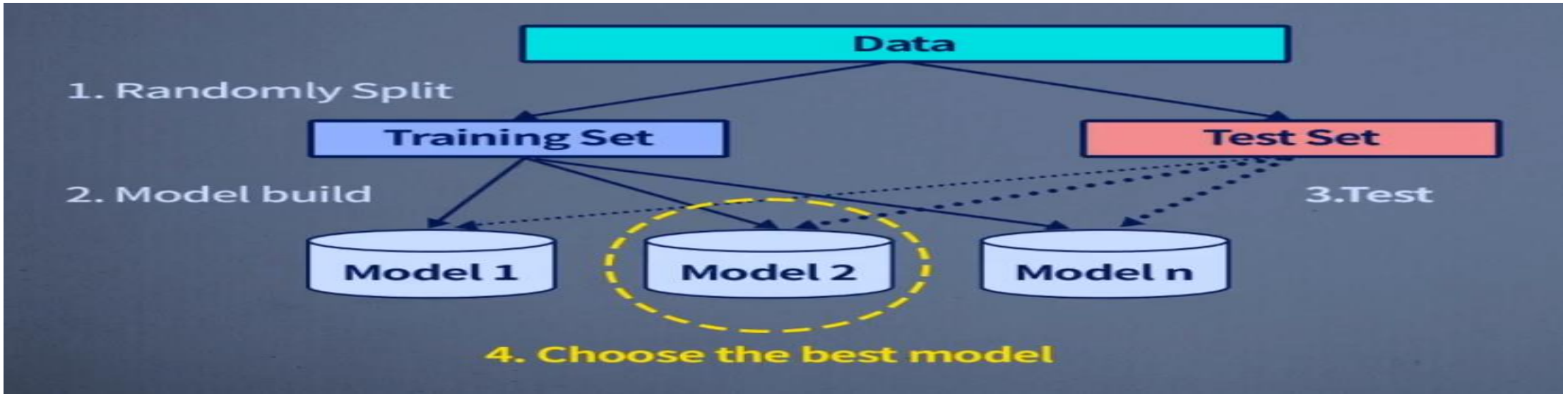
- 데이터들의 패턴을 대표할 수 있는 함수
- 어떤 Model을 선택해야 하는가?

Overfitting vs Generalization

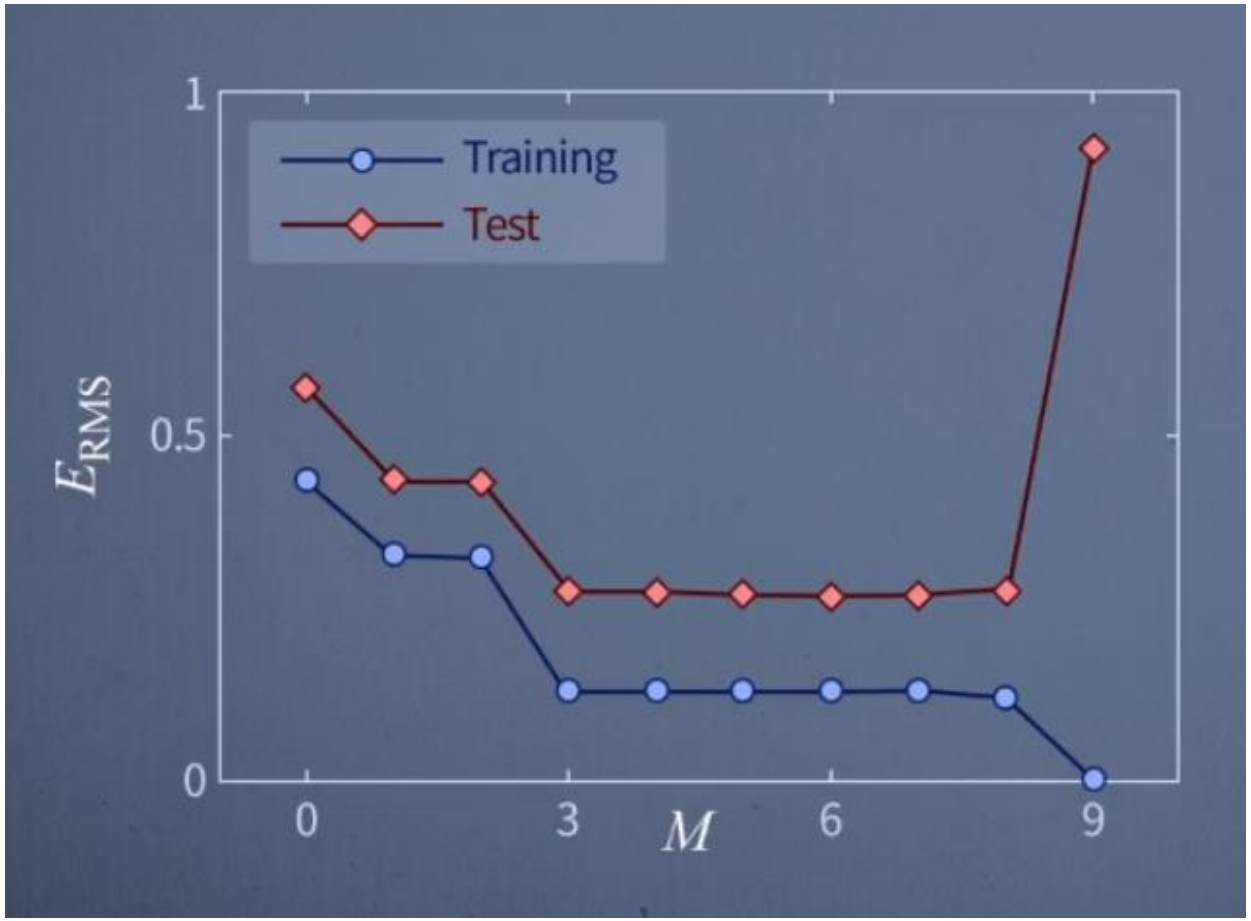
- 목표 : 주어진 data를 이용해서 앞으로 나올 unknown data 예측(주어진 데이터를 error없이 정확히 fitting [X])
- Overfitting
  - 모델의 차수(복잡도)가 올라가면 training data를 더 잘 학습 가능
  - 즉, error 최소화 가능
  - 하지만 prediction 정확도는 떨어질 수 있음
  - Generalization 필요

좋은 모델 선정1–Training and Test Set

- 주어진 Data를 training set과 test set으로 나눈다.(Random하게)
- Model을 생성한다.
- 모델을 Test Data를 통해 정확도를 검정한다.
- 정확도가 높은 모델을 선택한다.



일반적 상황



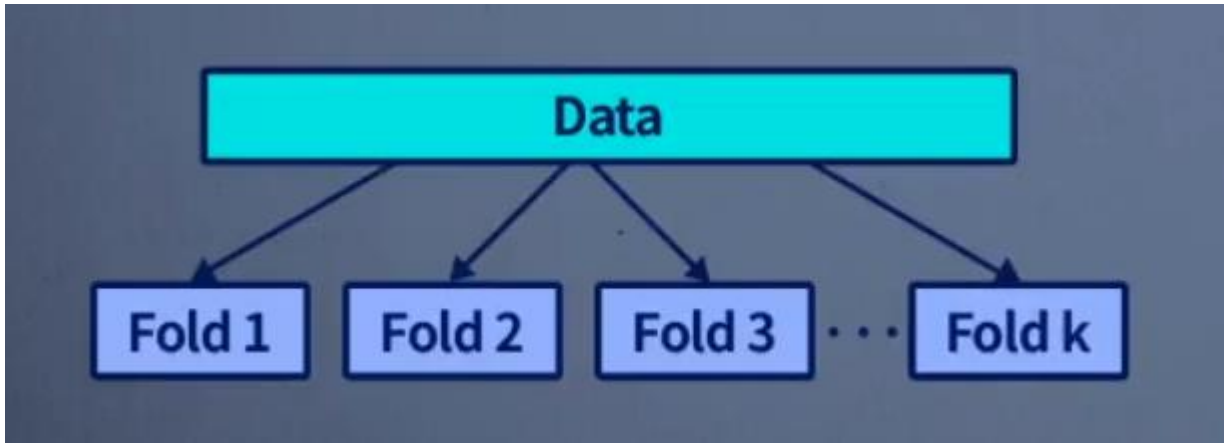
1. Traning Data 에러율 감소
2. Test Data 에러율 감소(Unknown Data이므로 에러율은 약간 높음)
3. 차수가 커지면 갑자기 에러율 커짐(무조건 차수가 높다고 좋은 모델은 아님)

장점과 단점

- 일반적으로 30%~50% 정도 Test data로 사용
- 장점
  - 간단하고 쉽다.
- 단점
  - Random하게 Dataset을 나누는 문제(같은 모델을 사용하더라도 데이터 셋에 따라 성능 차이가 날 수 있음)

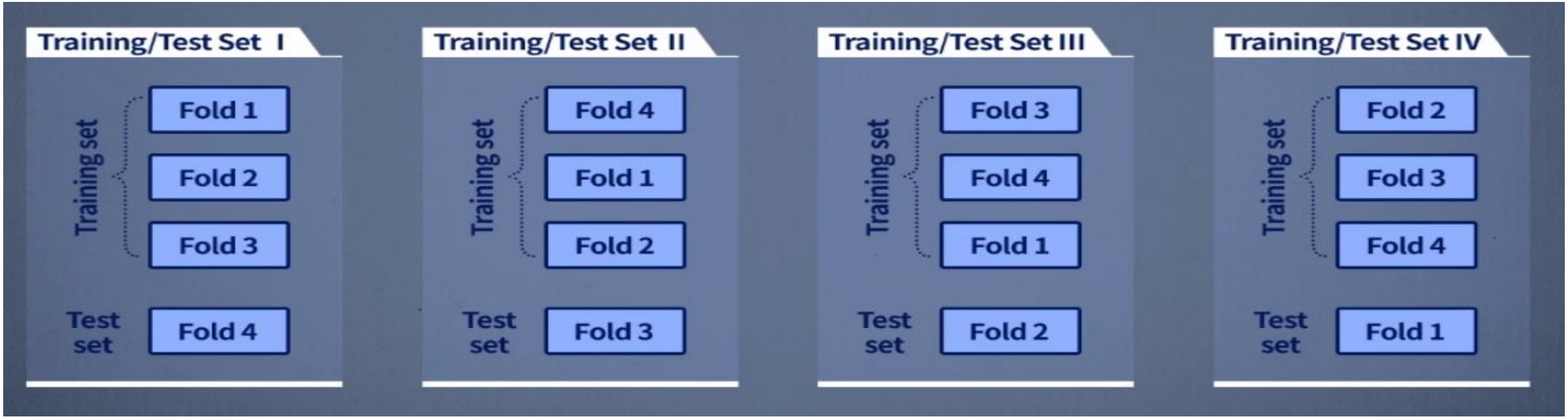
좋은 모델 선정2–Cross Validation

- Training and Test set 확장
- K-fold Cross Validation
  - 주어진 데이터를 k개의 fold로 나눈다. (fold = group)

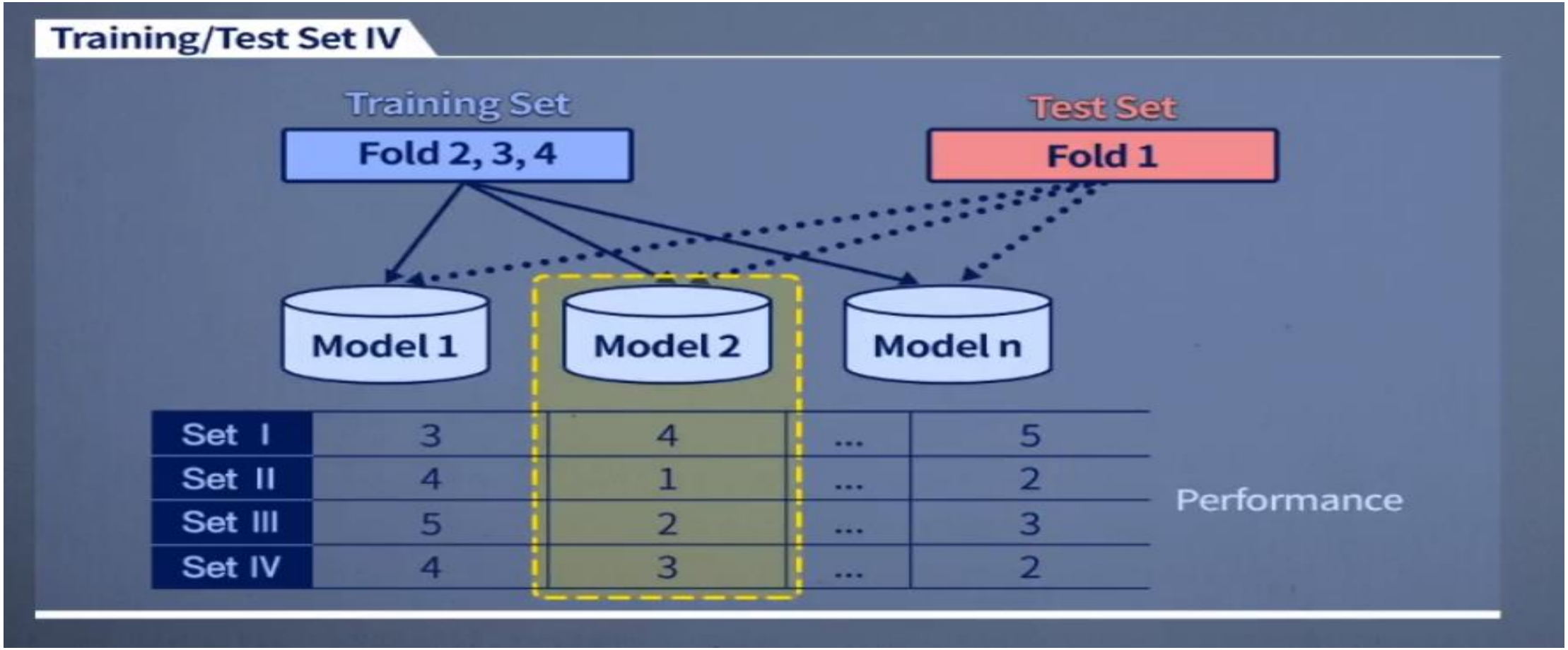


4 fold cross validation

- 평가 4개에 대한 평균값으로 모델의 성능 결정



4 fold cross validation



장점과 단점

- 장점
  - 모든 date를 Training과 Test에 사용 가능하다.
  - K번 model을 evaluation해서 평균값으로 평가, 작은 variance 평가
- 단점
  - 시간이 오래걸린다.