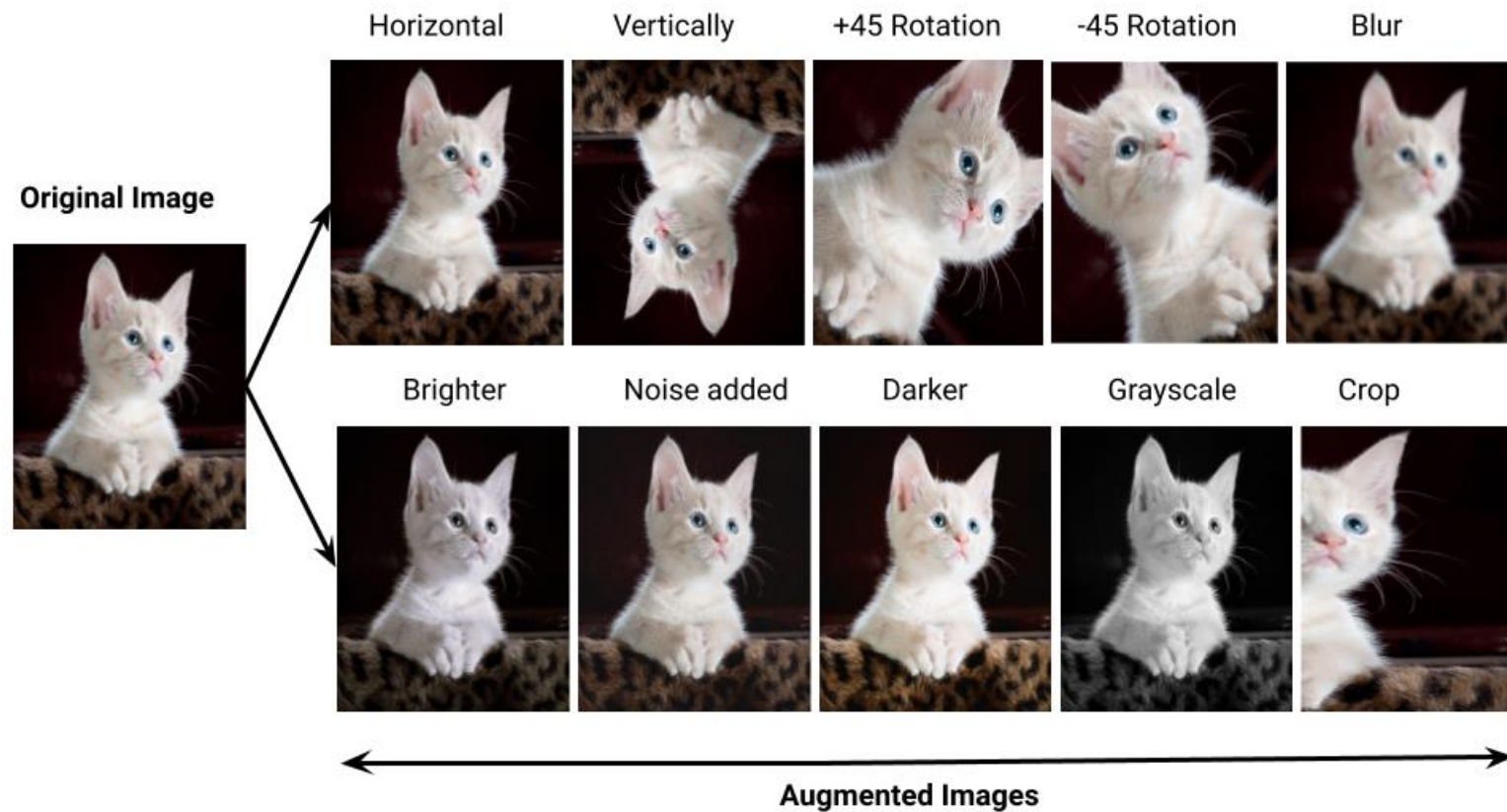


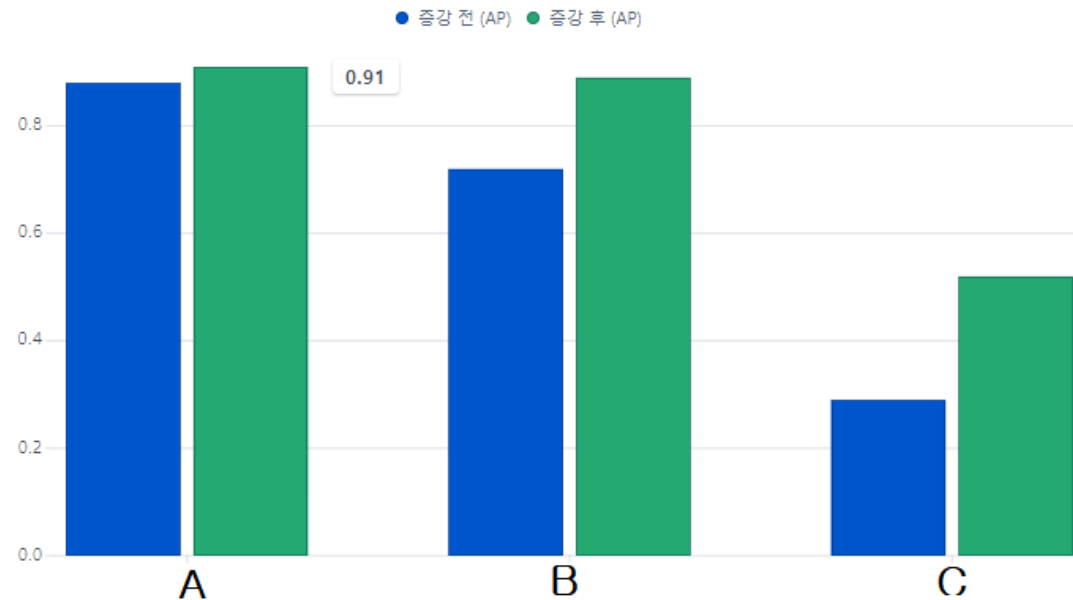
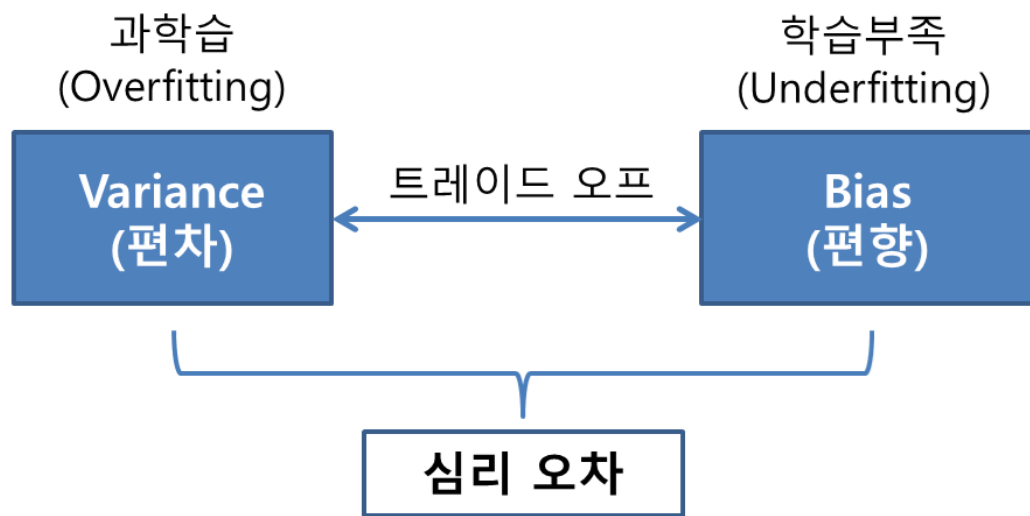
# AI 실전 4주차

어떻게 데이터를 늘려볼까?  
데이터 증강

# 데이터 증강?



# Why? 데이터 증강 필요



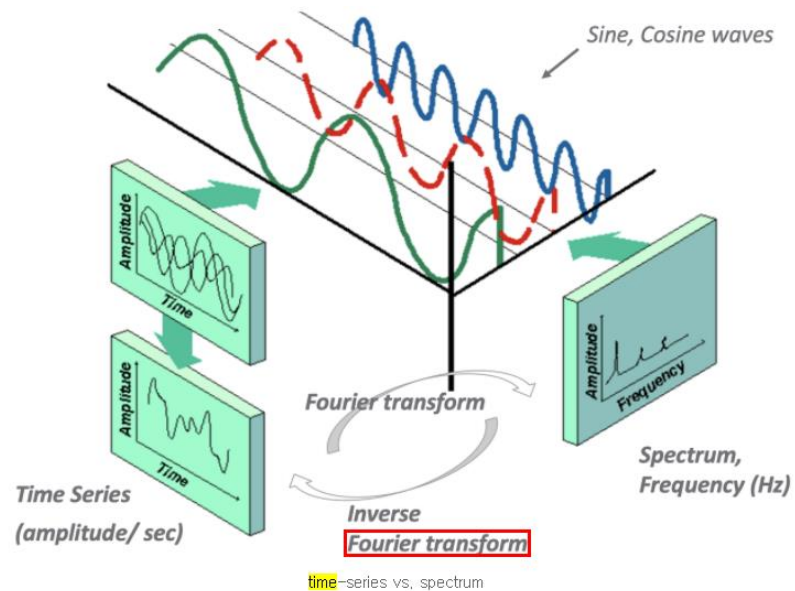
# 데이터 증강 종류(이미지)

## Base Augmentations



# 시계열 데이터 증강

- 데이터의 일정 시간 간격으로 기록됨.
- 시간의 종속성으로 인해 데이터를 뒤죽박죽 조합하면 안된다!



# 시계열 데이터 증강 방법

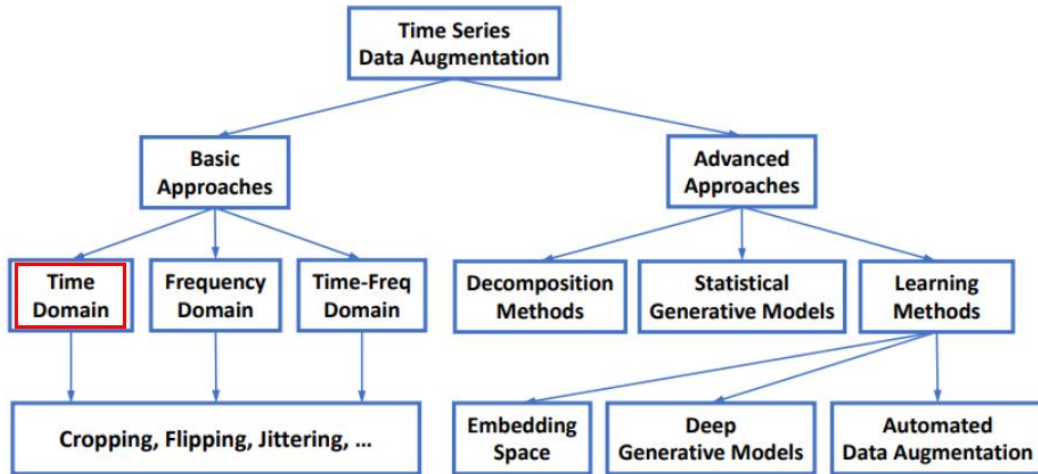


Figure 1: A taxonomy of time series data augmentation techniques.

- **Window cropping or slicing**
  - 원본 데이터에서 랜덤하게 연속적인 slice를 추출
  - 이미지를 랜덤하게 잘라 데이터를 증강하는 것과 유사
- **Window warping**
  - 랜덤한 구간을 선정하여 압축 or 확장하는 방법이다.
  - 원본 시계열의 전체 길이를 바꿀 수 있다.
- **Flipping**
  - 원본 데이터의 부호를 바꿔 새로운 시퀀스를 만드는 방법
  - 부호를 바꿔도 라벨 동일
- **Noise injection**
  - 라벨을 바꾸지 않은 채 원본 데이터에 노이즈를 삽입
  - 가우시안 노이즈(정규분포를 갖는 잡음) 또는
  - spike, step-like trend, and slope-like trend 등 적용
- **Label expansion**
  - 이상치 감지를 위한 증강 기법이다.
  - 거리와 값 거리가 이상치와 가까운 이상치로 분류하여 라벨을 확장

# 자연어 데이터 증강

Operation	Sentence
None	A sad, superior human comedy played out on the back roads of life.
SR	A <i>lamentable</i> , superior human comedy played out on the <i>backward</i> road of life.
RI	A sad, superior human comedy played out on <i>funniness</i> the back roads of life.
RS	A sad, superior human comedy played out on <i>roads</i> back <i>the</i> of life.
RD	A sad, superior human out on the roads of life.

Table 1: Sentences generated using EDA. SR: synonym replacement. RI: random insertion. RS: random swap. RD: random deletion.

EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks

- Synonym Replacement, SR
  - 문장에서 랜덤으로 stop words가 아닌 n 개의 단어들을 선택해 임의로 선택한 동의어들 중 하나로 바꾸는 기법.
- Random Insertion, RI
  - 문장 내에서 stop word를 제외한 나머지 단어들 중에서, 랜덤으로 선택한 단어의 동의어를 임의로 정한다. 그리고 동의어를 문장 내 임의의 자리에 넣는걸 n번 반복한다.
- Random Swap, RS
  - 무작위로 문장 내에서 두 단어를 선택하고 위치를 바꾼다. 이것도 n번 반복
- Random Deletion, RD
  - 확률 p를 통해 문장 내에 있는 각 단어들을 랜덤하게 삭제한다.

# 자연어 처리 증강 성능 비교

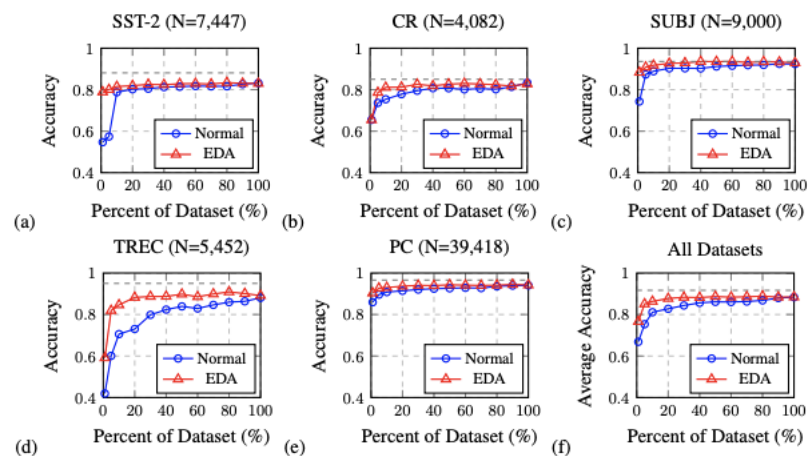


Figure 1: Performance on benchmark text classification tasks with and without EDA, for various dataset sizes used for training. For reference, the dotted grey line indicates best performances from Kim (2014) for SST-2, CR, SUBJ, and TREC, and Ganapathibhotla (2008) for PC.

Model	Training Set Size			
	500	2,000	5,000	full set
RNN	75.3	83.7	86.1	87.4
+EDA	79.1	84.4	87.3	88.3
CNN	78.6	85.6	87.7	88.3
+EDA	80.7	86.4	88.3	88.8
Average	76.9	84.6	86.9	87.8
+EDA	79.9	85.4	87.8	<b>88.6</b>

Table 2: Average performances (%) across five text classification tasks for models with and without EDA on different training set sizes.