



deeplearning.ai

Mismatched training  
and dev/test data

---

Addressing data  
mismatch

# Addressing data mismatch

- • Carry out manual error analysis to try to understand difference between training and dev/test sets

E.g. noisy - car noise, mis-recognizing street numbers

- • Make training data more similar; or collect more data similar to dev/test sets

E.g. simulate noisy in-car data, 사방이 소리 있는 컴퓨터를 더 모으자.

training data를 더 dev set이랑 비슷하게 만들기 위해 할 수 있는 일이 뭘까?  
-> artificial data synthesis

# Artificial data synthesis

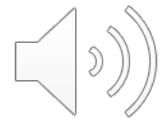
만약 error analysis 결과가 이 과정이 필요하다고 한다면 차 타고 밖에 나가서 수백 수천 시간의 새로운 트레이닝 데이터를 수집하는 것보다는 훨씬 reasonable한 프로세스가 될 것이다.



+



=

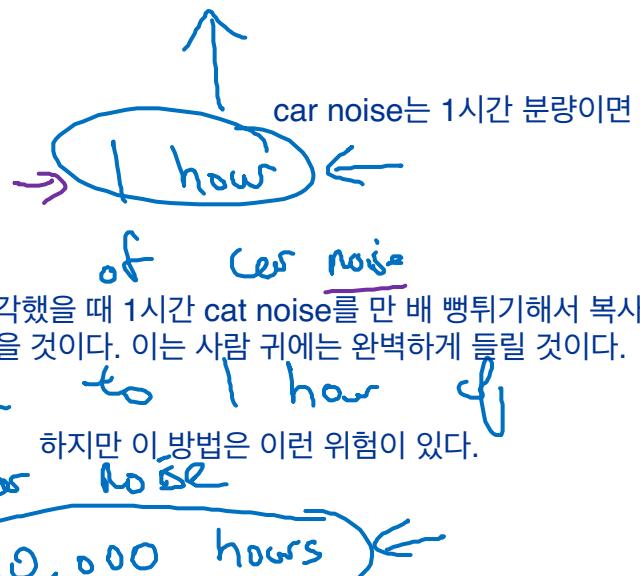


“The quick brown ←  
fox jumps  
over the lazy dog.”

10,000 hours

general speech data는 10,000 시간 분량이고

Car noise



Synthesized  
in-car audio

Set of all audio in car

이 space의 매우 작은 subset만 시뮬레이팅하는 거고 이 작은 subset에서만 synthesizing하는 것일지도 모른다.

Andrew Ng

# Artificial data synthesis

## Car recognition:

힘들게 실제 차 사진 수집할 것 없이 그냥 컴퓨터로 만들어서 쓰면 안되나?

안된다. 왜냐하면 사람 눈으로 보기에는 문제없어 보일지라도 앞의 경우와 마찬가지로 극히 일부의 subset 많 합성하는 것일 수 있고 그렇다면 이 작은 subset에 오버피팅 될 수도 있을 것이다.



N<sup>20</sup> cars

