deeplearning.ai

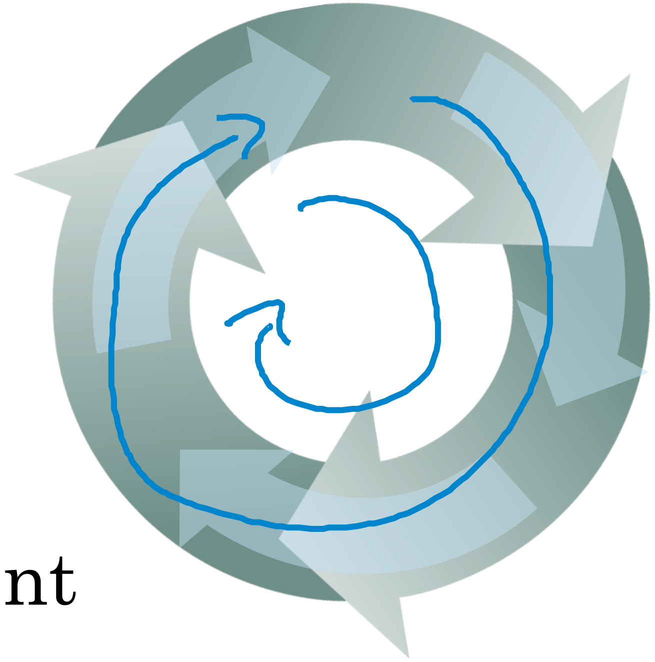
Setting up your ML application

Train/dev/test sets

Applied ML is a highly iterative process

이렇게 다양한 하이퍼파라미터가 있는데

- # layers
- # hidden units
- learning rates
- activation functions
- …



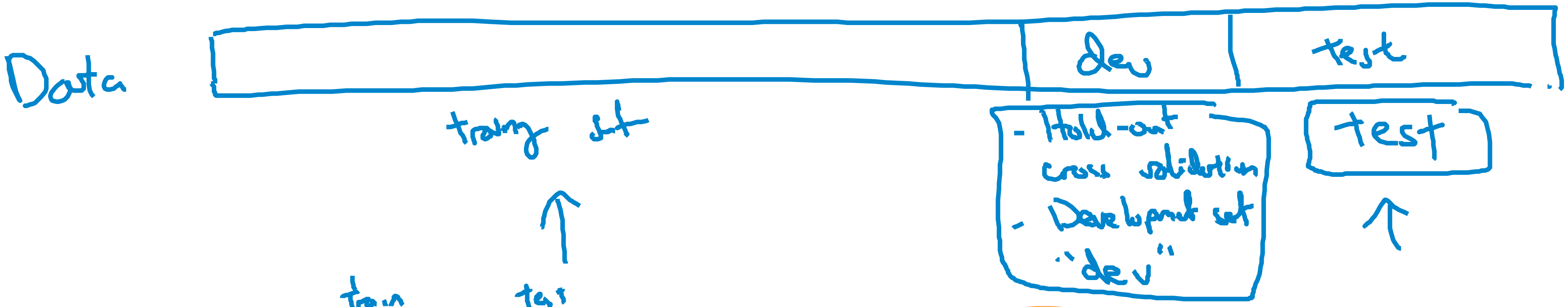
그래서 아무리 경험이 많아도 한방에 최적의
하이퍼 파라미터를 알아내기는 힘들고
반복적으로 작업을 해야함

한 도메인에서의 직관이 다른곳에서는 통하지 않는 경우 많다.

NLP, Vision, Speech, Structural data

Ads   search   Security   logistic ....

Train/dev/test sets



train test

Previous era: 70/30%   60/20/20%

previous era

100 - 1000 - 10000

Big data: 1,000,000   10,000   10,000

Big data era

98/1/1%

99.5/ ~~.25~~ .4 / ~~.25~~ .1%

전통적인 방법론에서 사용하는것처럼 10~20%
씩이나 dev/test set에 사용할 필요는 없음

Andrew Ng

Mismatched train/test distribution

Cats

Training set:
Cat pictures from
webpages

고품질 vs. 저품질

이 둘의
distribution 종종
다르다

Dev/test sets:
Cat pictures from
users using your app

→ Make sure dev and test come from same distribution.

가이드라인: 트레이닝 데이터가 많은게 중요하므로 크롤링같은걸 할텐데
이때에도 dev-test distribution 같도록 하는게 중요하다.

"test"

train / dev

기존 머신러닝에서는 test set을
dev set으로 부르기도 한다/

train / test

→ Train / dev

Not having a test set might be okay. (Only dev set.)