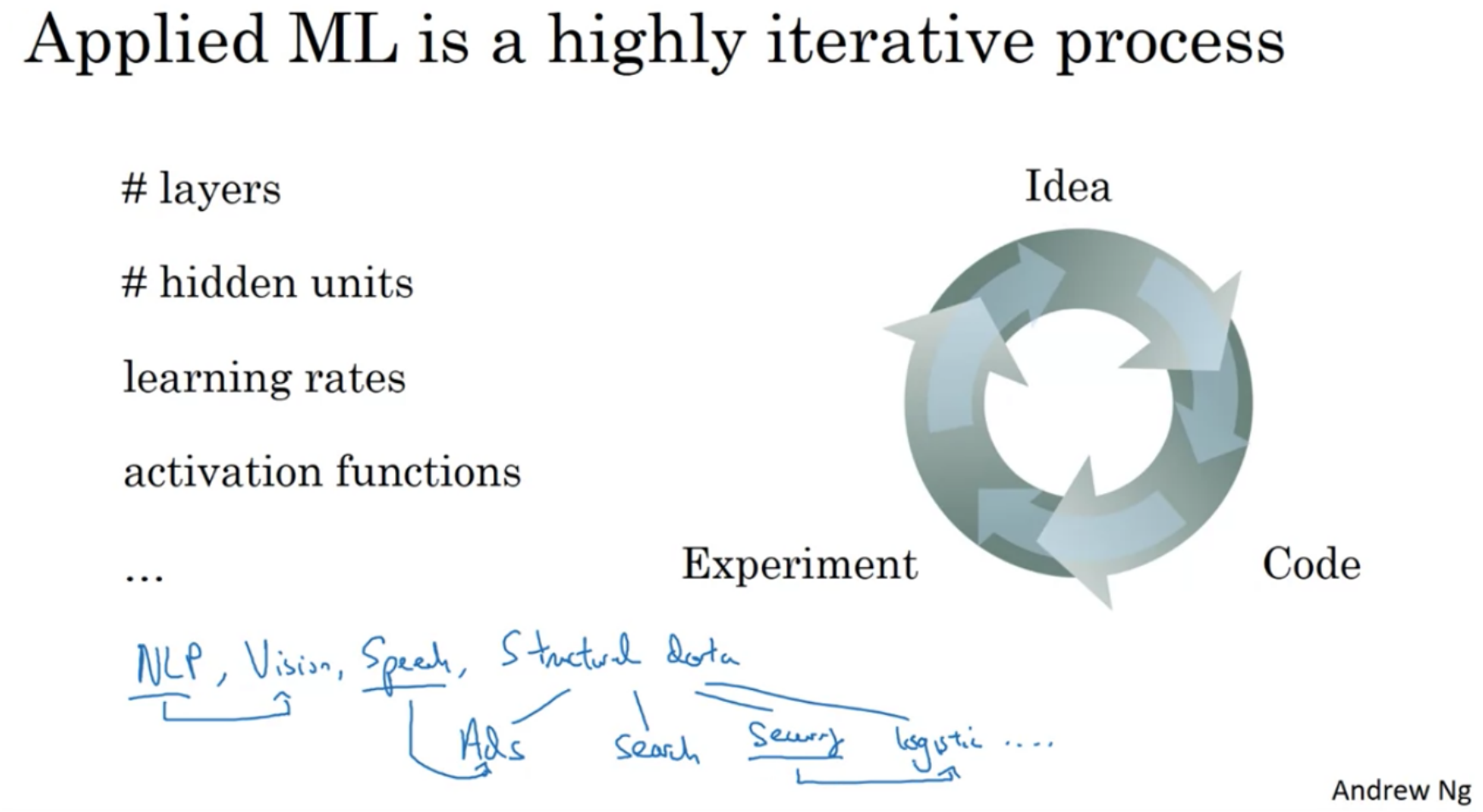
Train/Dev/Test sets

Tags Train/Dev/Test

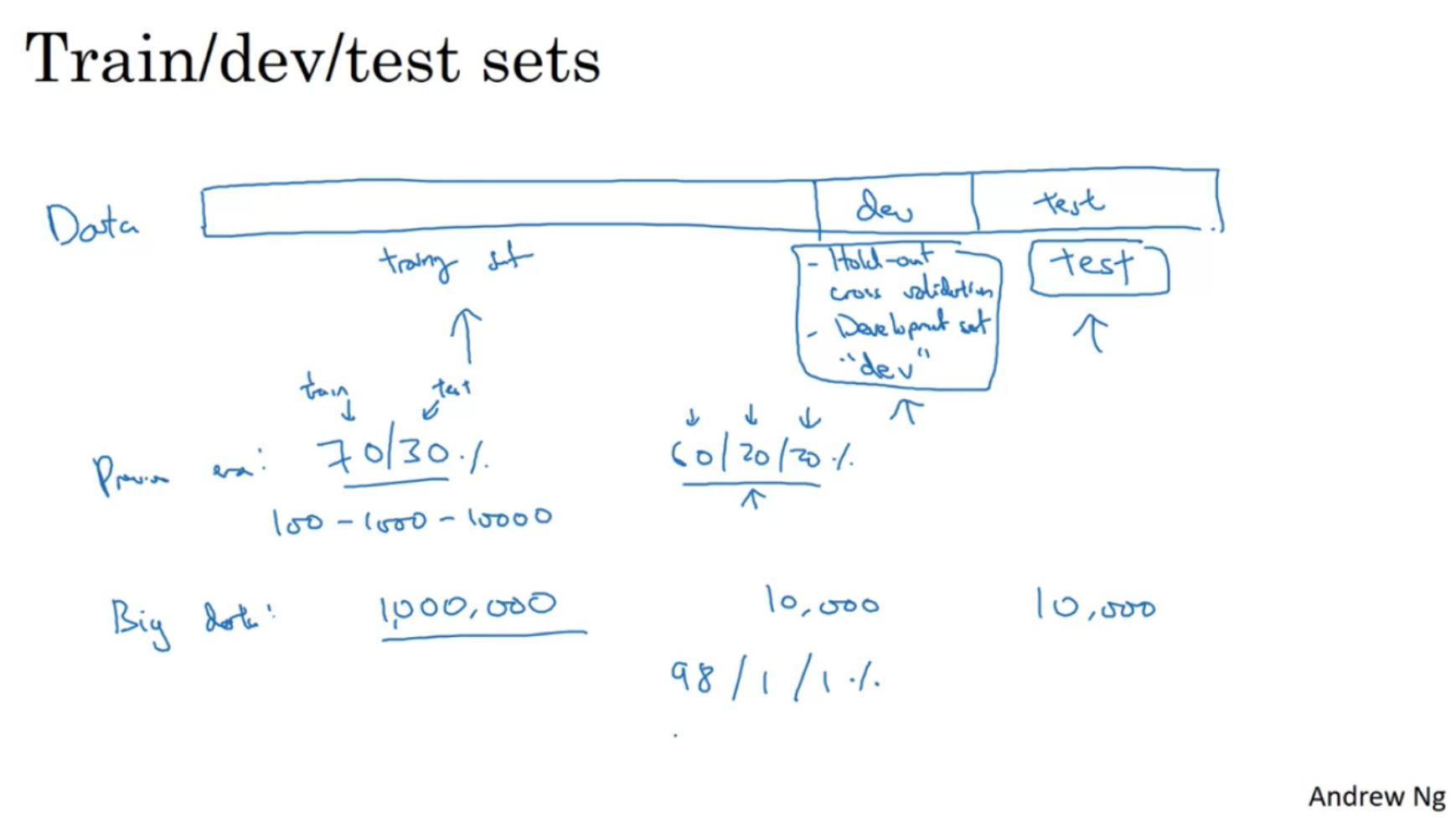
Applied ML is highly iterative process

Mismatched train/test distribution

Applied ML is highly iterative process



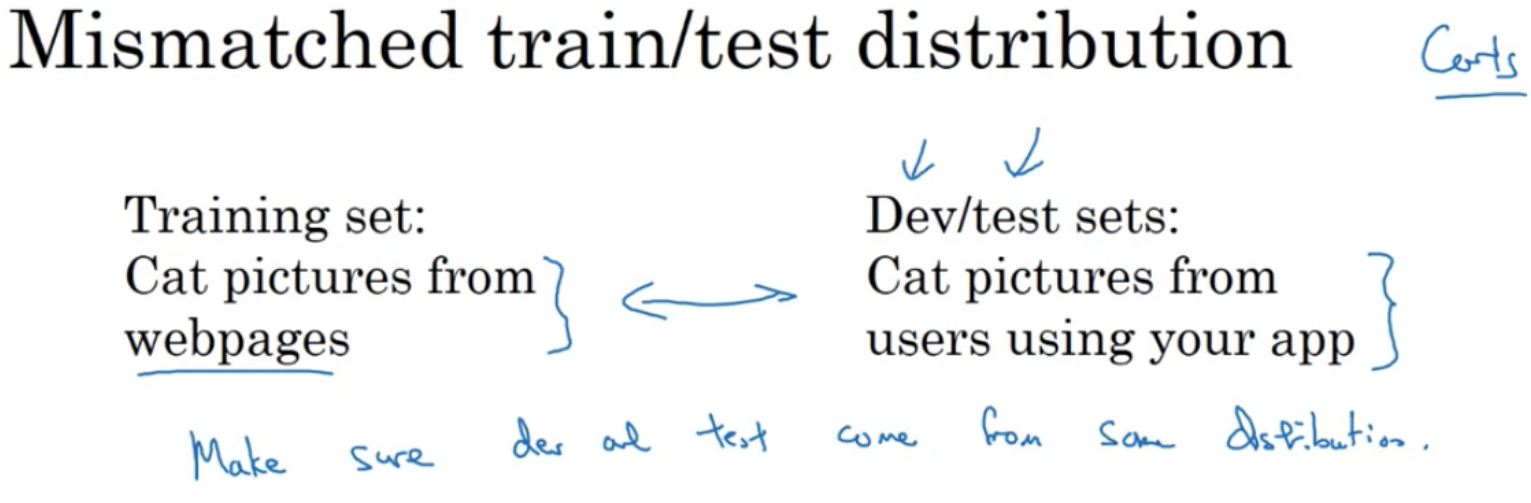
> So one of the things that determine how quickly you can make progress is how efficiently `you can go around this cycle`. And setting up your data sets well in terms of your train, development and test sets can make you much more efficient at that.



- 100만개가 넘는 데이터를 갖고 있다면 train/dev/test 비율을 어떻게 정해야할까?
- ⇒ 1%의 dev set, 1%의 test set으로도 충분하다

> So just to recap, when setting up your machine learning problem, I'll often set it up into a train, dev and test sets, and if you have a relatively small dataset, these traditional ratios might be okay. But if you have a much larger data set, it's also fine to set your dev and test sets to be much smaller than your 20% or even 10% of your data.

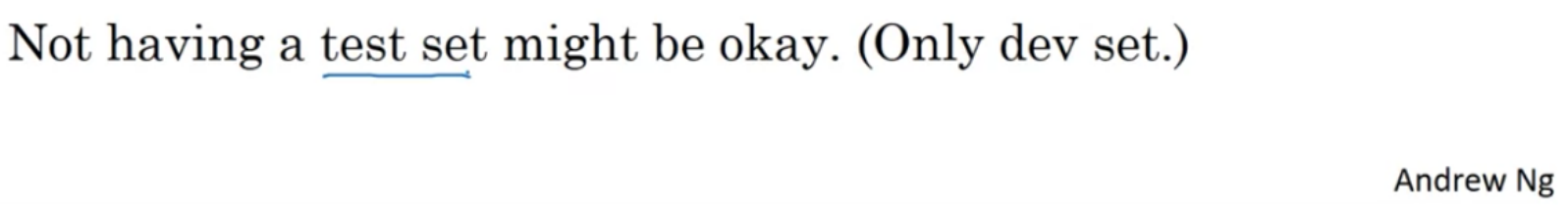
Mismatched train/test distribution



💡 Make sure dev and test set come from same distribution

> 한 가지 가정을 해보자.
> 고양이 분류기를 만든다고 하자.
> 분류기의 학습을 위해 학습 데이터는 웹에서 고화질로 크롤링했다. 반면 App을 통해 수집되는 데이터는 화질이 좋지 않을 수도 있고 블러리 할 수도 있다.
> 이경우, 어떤 문제점이 발생할 수 있을까?
>
> Train set과 Dev/Test Set의 분포가 다르기에 좋은 결과로 이어지지 않을 수 있다

> Test Set을 갖지 않아도 괜찮다?!



> And this is actually okay practice if you don't need a completely unbiased estimate of the performance of your algorithm.

⇒ *Question. What it means?!*