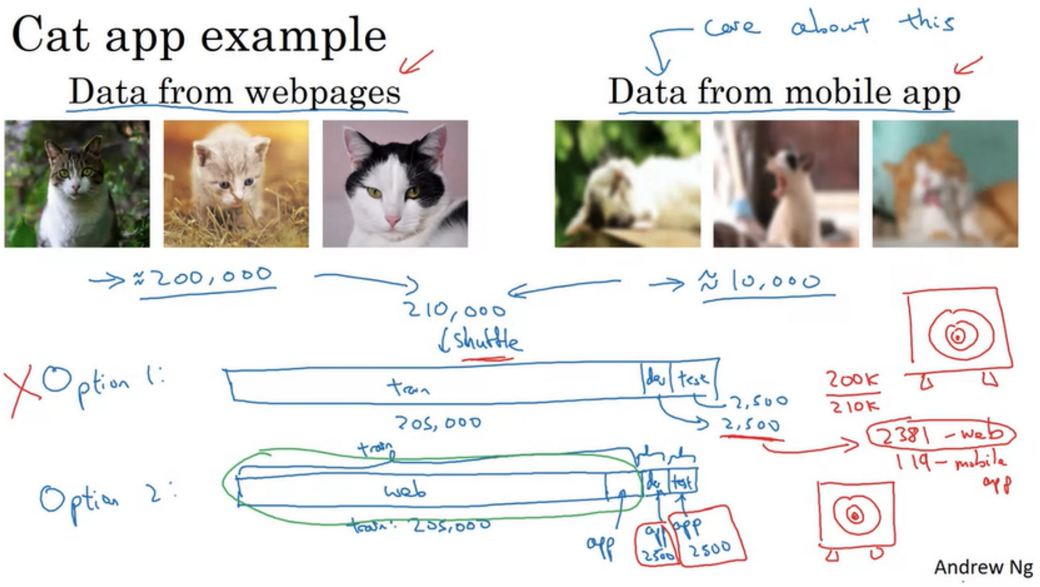
Mismatched training and dev/test set

Tuesday, September 1, 2020 9:24 PM

① Training and testing on different distributions



→ Prioritize the data you care for in the dev / test set

② Bias and Variance with mismatched data distributions

→ Since Training and dev set are different, we can't evaluate bias and variance

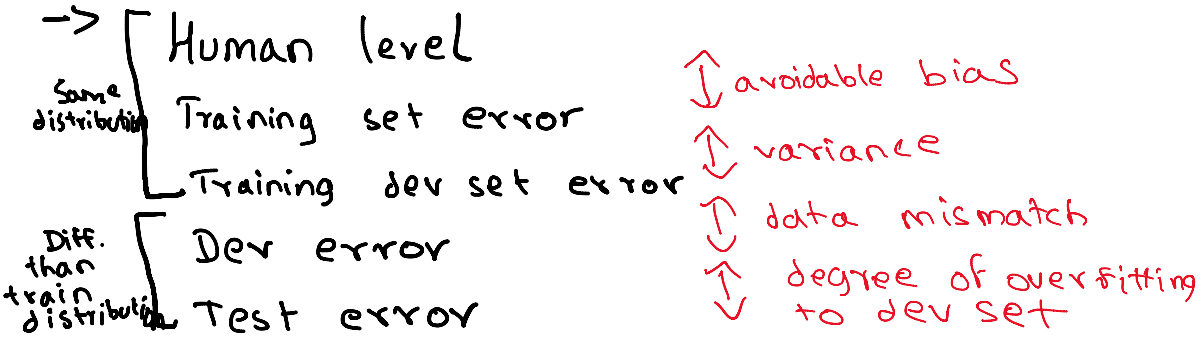
→ We define Training dev-set: Same distribution as training set, but not used for training

→ eg:

Training error	1%	↕ variance problem	1%	
Training-dev error	9%		1.5%	↕ data mismatch problem
Dev error	10%		10%	

eg:

Human error	0%	↕ avoidable bias problem	0%	↕ avoidable bias problem
Training error	10%		10%	
Training-dev error	11%		11%	↕ Data mismatch problem
Dev error	12%		20%	



- error doesn't necessarily go up as you go from Training → Test error

③ Addressing Data Mismatch

- Carry out manual error analysis to try to understand difference between training and dev/test sets.
- Make training data more similar; or collect more data similar to dev/test sets (eg: Artificial data synthesis (adding noise))
  - → Artificial data synthesis might lead to overfitting of model to tiny subset of the data that we can simulate