deeplearning.ai

Mismatched training and dev/test data

Training and testing on different distributions

Cat app example

Data from webpages



Data from mobile app

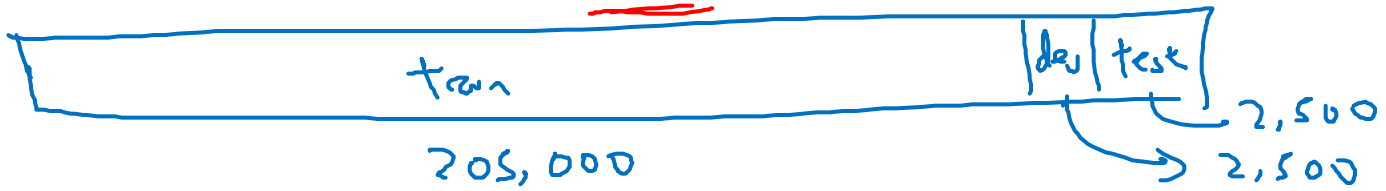


care about this

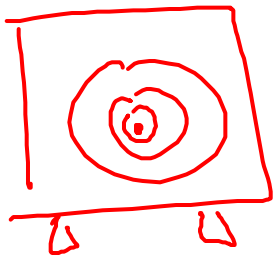
≈200,000 → 210,000 ← ≈10,000

(shuffle

Option 1:



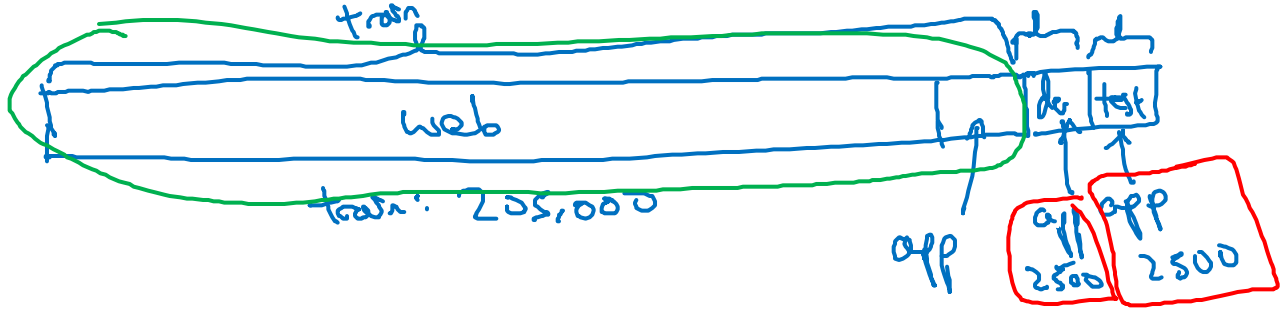
200K / 210K

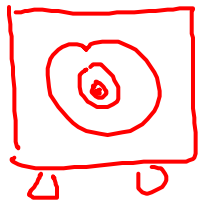


2381 - web

119 - mobile app

Option 2:





Speech recognition example

Speech activated rearview mirror

Training

- Purchased data
- Smart speaker control
- Voice keyboard
- ...

500,000 utterances

Dev/test

Speech activated rearview mirror

20,000

10K 5K 5K D T

train

500K

D T

510K

10K mirror

5K 5K

deeplearning.ai

Mismatched training and dev/test data

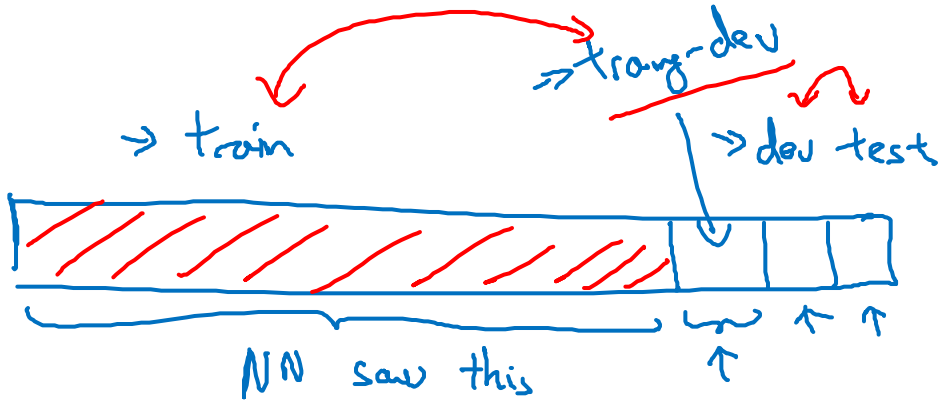
Bias and Variance with mismatched data distributions

Cat classifier example

Assume humans get ≈ 0% error.

Training error 1%
Dev error 10%

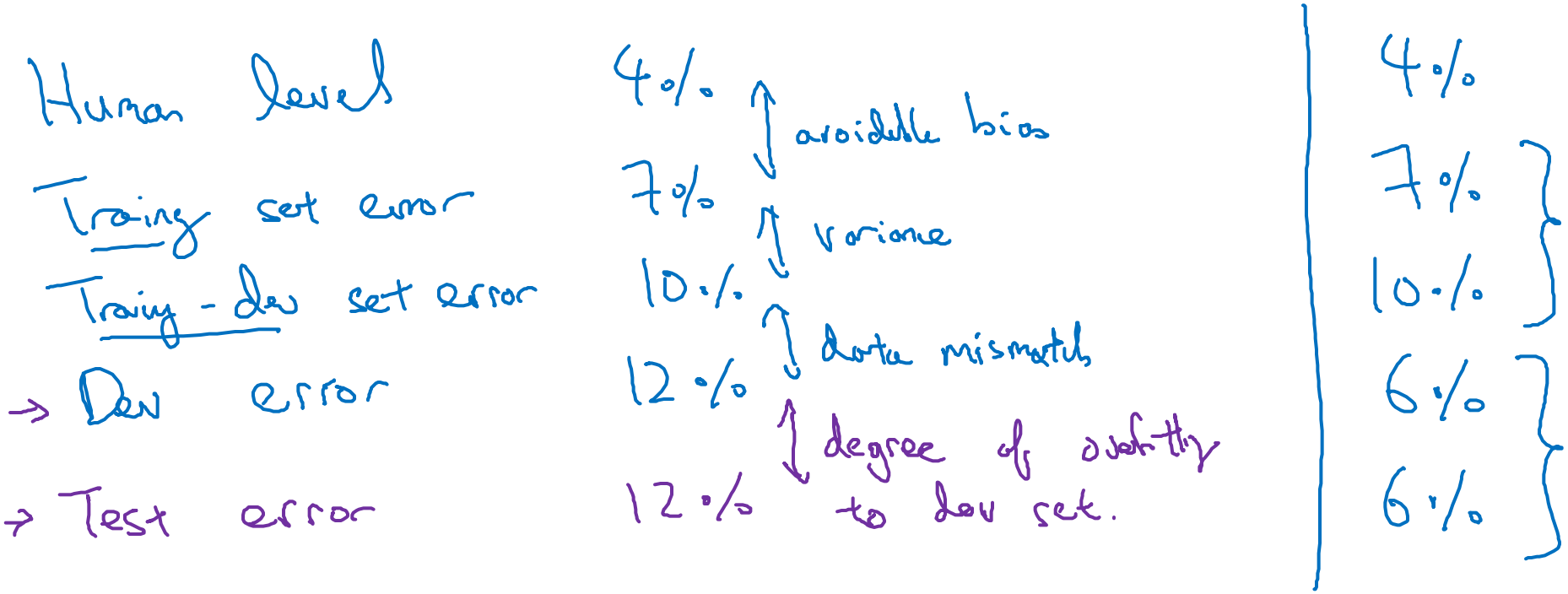
Training-dev set: Same distribution as training set, but not used for training



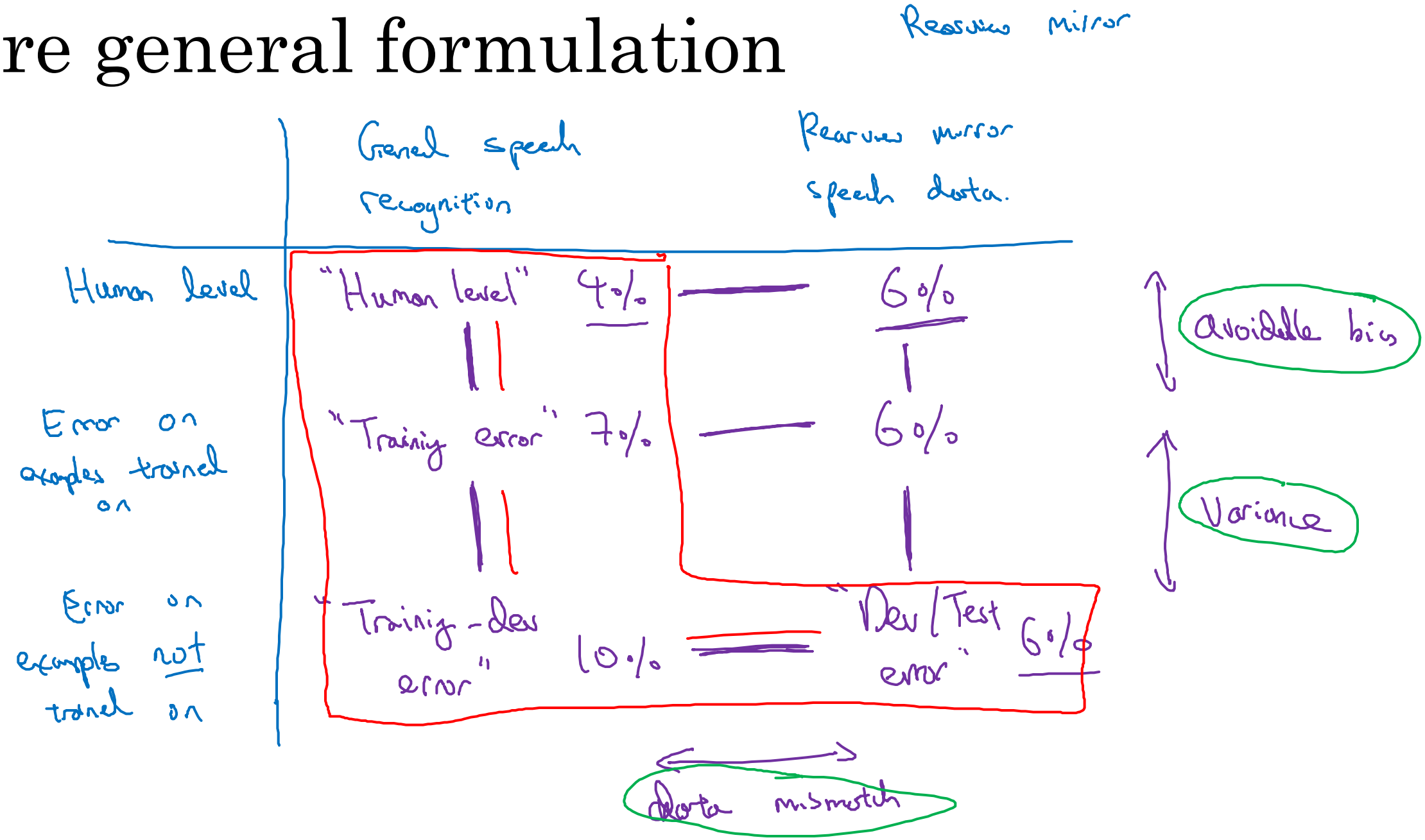
Training error 1% → 1% (variance)
→ Train-dev error 9% → 1.5% (data mismatch)
→ Dev error 10% → 10%

	Variance	
Human error	0%	Avoidable bias
Training error	10%	Avoidable bias / Variance
Train-dev error	11%	Variance / Data mismatch
Dev error	12%	
	Bias	Bias + Data mismatch

Bias/variance on mismatched training and dev/test sets



More general formulation



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 street numbers

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

E.g. Simulate noisy in-car data

Artificial data synthesis

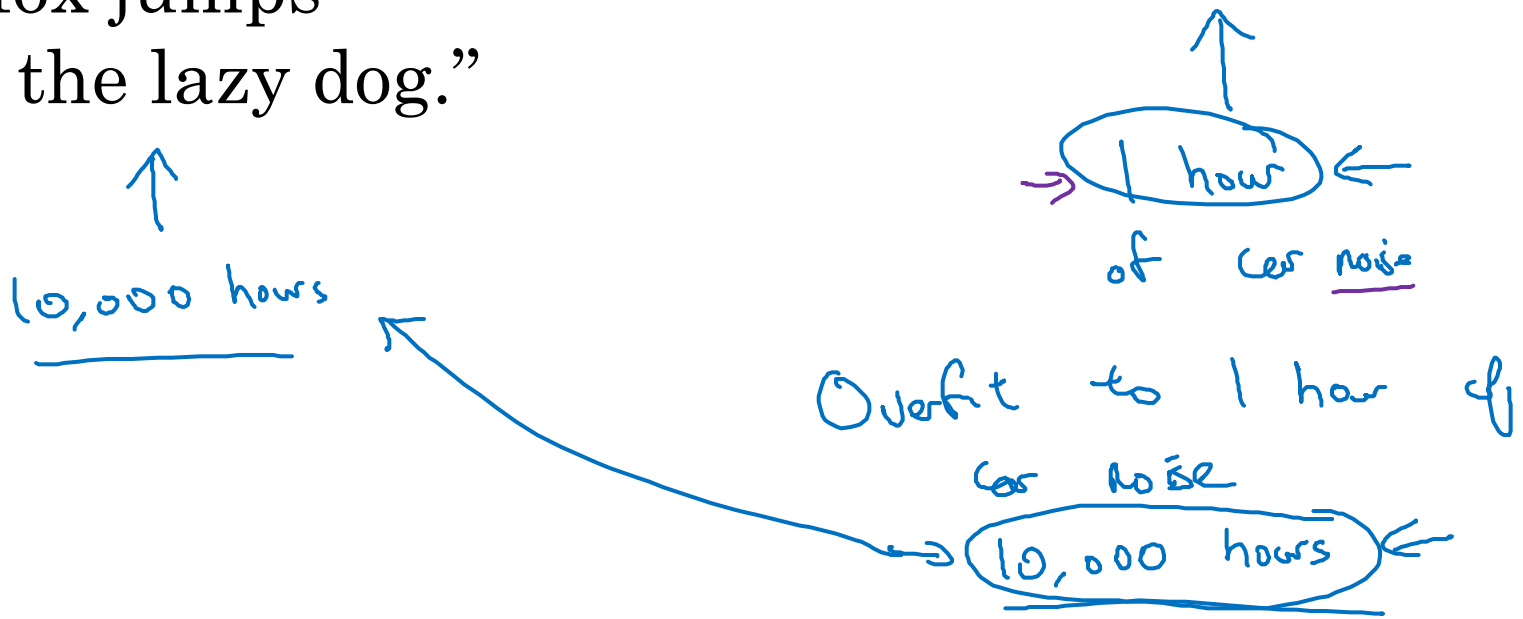
+

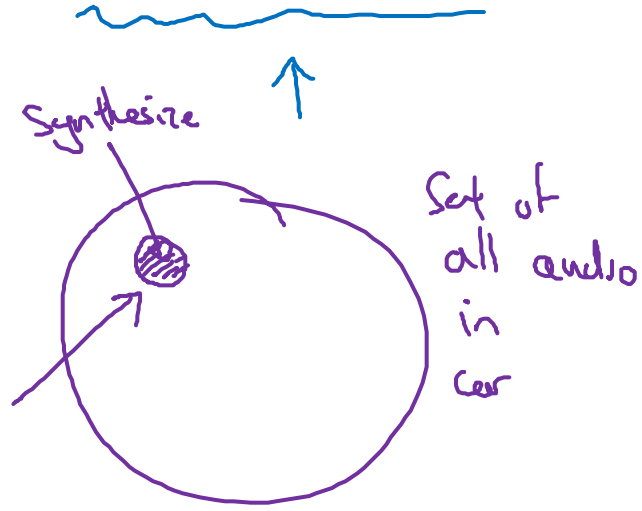
=

"The quick brown fox jumps over the lazy dog."

Car noise

Synthesized in-car audio





Andrew Ng

Artificial data synthesis

Car recognition:

≈ 20 cars