Q Should you Always use all the data you have

Mismatched training and dev/test data

Bias and Variance with mismatched data distributions

deeplearning.ai

Cat classifier example

- Estimating Absolute bias & variance helps when you want to find what to focus on ↓, bias or variance?
- But, the way we calculate bias & variance is diff when training dist is diff compared to dev/test dist

Assume humans get ≈ 0% error.
(Bayes error)

Training error 1%
Dev error 10%

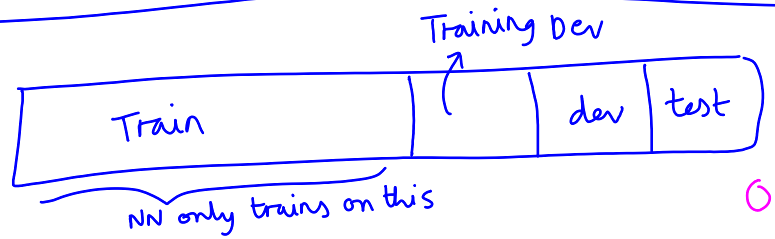
→ If dev data came from the same dist as training data, we would say we have large variance problem & work to ↓ variance

- But what if it didnt come from same dist?

eg training data was very easy → high Res Images
∴ error = 1%, dev set is blurry Images ∴ 10%
⇒ its not a variance problem, its a data mismatch problem

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

New data set (dont explicitly train the NN on this)



Same dist

			0% (Human)	0%
			↕ large	↕ large
Training error :	1%	1%	10%	10%
	↕ large			
Training Dev error :	9%	1.5%	11%	11%
		↕ large		↕ large
Dev error :	10%	10%	12%	20%
	Variance problem	Data mismatch problem	Avoidable bias	2 problems ① Avoidable bias ② Data mismatch

Bias/variance on mismatched training and dev/test sets

Human error : 4%
↕ Avoidable bias
Training set error : 7%
↕ Variance
Training dev set error : 10%
↕ data mismatch
Dev error : 12%
↕ degree of overfitting to dev set
Test error : 12%

same dist (Training set error, Training dev set error)

same dist (Dev error, Test error)

4%
7%
10%
6%
6%

Dev/Test set can have lesser error compared to Train sets
⇒ Training data was more difficult than dev/test set

More general formulation

Speech Activated Rear view mirror

purchased data, samples from Amzn Echo etc.

	General Speech Recognition data samples	Rear view mirror speech data
Human level	"Human level" 4%	6%
Error on examples the NN has trained on	"Training error" 7%	6%
Error on examples the NN has NOT trained on	"Training Dev set error" 10%	"Dev/Test set error" 6%

Avoidable bias

Variance

data mismatch

How to get this? Make your Audience use the App

How to get this? Train NN on samples from the Rear view App

⟹ If we get 6% Human & 6% training from NN, we're already at human level performance

We can handle avoidable bias & variance, How to handle data mis match?

Next slide