Orthogonalization

Fit training set well on cost function
↓
Fit dev set well on cost function
↓
Fit test set well on cost function
↓
Performs well in real world

Bigger network
Adam
...
Early Stopping X
Regularization
Bigger Train set
Bigger Dev Set
Change Dev Set or Cost Function

Setting up Goal

Single number evaluation metric

e.g.

Model	Precision	Recall
A	95%	90%
B	98%	85%

⇒ F1 Score = $\frac{2}{\frac{1}{P}+\frac{1}{R}}$ ("Harmonic Mean")

Satisficing and optimizing metric

e.g.

Model	Accuracy	Running Time
A	90%	80 ms
B	92%	95 ms
C	95%	1500 ms

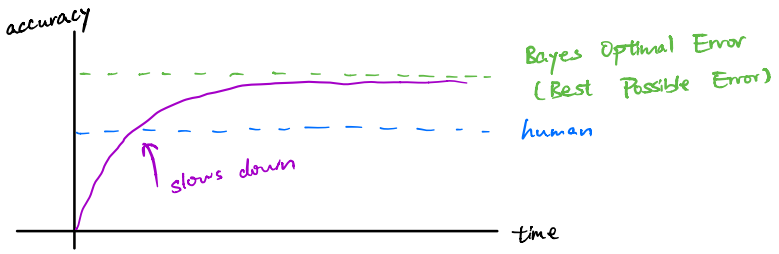
⇒ maximize accuracy — optimizing metric
subject to running time ≤ 100 ms — satisficing metric

Train / Dev / Test Set

e.g. Market A, B, C → Dev
Market D, E, F → Test
This is NOT good!

⇒ Choose a dev set and a test set to reflect data you expect to get in the future

Compare to Human-Level Performance



Why ?
- Get labeled data from humans
- Gain insight from manual error analysis
- Better analysis of bias / variance

Avoidable Bias

Humans	1%	7.5%	Proxy of Bayes Error
Training Error	8%	8%	"Avoidable Bias"
Dev Error	10%	10%	"variance"

Understanding Human-Level Performance

e.g. Medical image classification
=> Bayes Error from "team of experienced doctors"

e.g. Team of humans		0.5%	0.5%
One human	"Avoidable Bias"	1%	1%
Training Error		0.6%	0.3%
Dev Error	"Var"	0.8%	0.4%

e.g. ML usually does better in structured data
(product recommendation ; loan approvals)

Improve model's performance

Human
↕ Avoidable Bias
Training Error
- Train Bigger model
- Train Longer / Better optimization algo
- NN architecture / hyperparameters search

↕ Variance
Dev Error
- more data
- regularization
- NN architecture / hyperparameters search