☆ Single Number Evaluation Metrics:

Recall: Actual identification of true subject
(identify cat as a cat)

Classifier	Precision	Recall	F1 score
A	95%	90%	92.4%
B	98%	85%	91%

F1 score : Avg. of P & R.

$$\frac{2}{\frac{1}{P}+\frac{1}{R}}$$ (Humonic mean)

Single evaluation metrics tells you clearly which classifier is better.

→ It will speed up iterating process

Average can also be used.

✡ Satisficing & Optimizing Metric:

Classifier	Accuracy	Running Time
A	90.1.	80ms
B	92.1.	95 ms
C	95.1.	1500 ms

Normal: accuracy

New: maximize accuracy
subject to running time ≤100

Optimise → accuracy
Satisfying → Running time

You have:
out of
N metrix: Optimise 1 matrix
N-1 → Satisfying var.

Wakewords | Trigger World

Taking above matrix as example,
Cost = accuracy − 0.5 x Running t

Now you can maximize cost.

In wake words: e.g. maximize accuracy
s.t. ≤1 false positive every 24 hr.
satisfying matrix → optimizing matrix

How to set up your dev / test set.

aka hold out cross validation set

- Your dev and test set should come from same distribution of data.

- using dev test you can iterate very quickly.

e.g. opti. on dev set on loan approvals for medium income zip codes.
X → y (repay loan?)
Tested on low income zip code.
At the performed bad due to bad distribution & bias.

Choose a dw / test set to reflect data you expect to get in the future and consider important to do well on.

Size of dw / test set:

Old way:

70% train | 30% test

60% train | 20% dw | 20% test

} good for lower data 10,000 pts.

for huge data sets:

98% train | 1% dw | 1% test

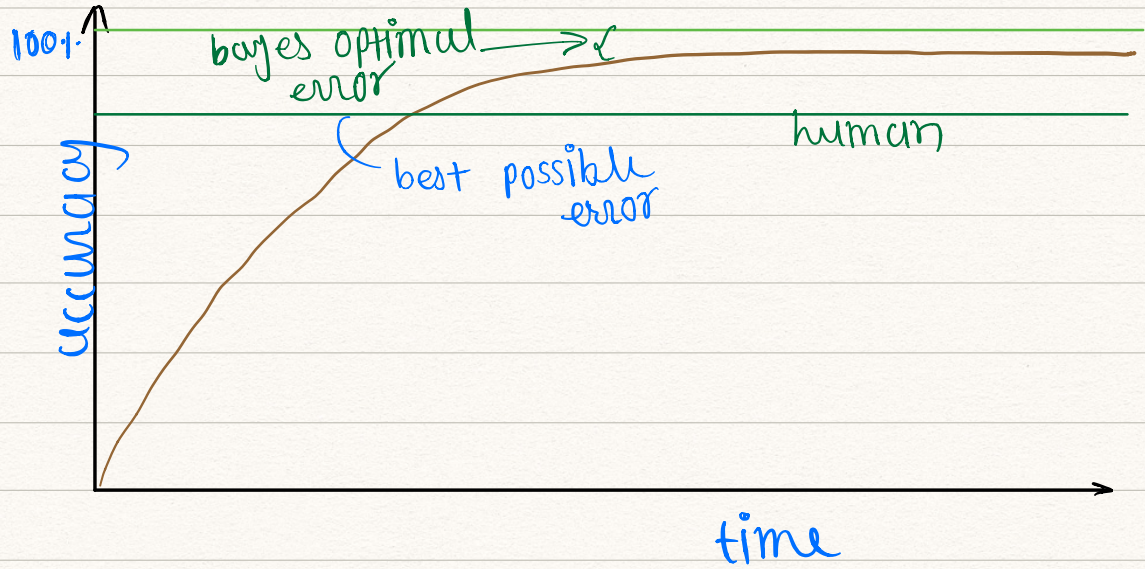
big enough to give high confidence in the overall performance of your system.

When to dev / test sets and metrics:

metrix + dev prefer A
you, user " B } need to change metrix

apply weight to reduce unaccepted input.

☆ Comparing to human level performance:



- You can not surpass bayes optimal error.
- There is not a lot of room for improvement once human level crossed.

Why compare to human level performance

Humans are quite good at a lot of tasks.

- get labeled data from humans
- gain insights from manual error analysis: Why did a person get this right?
- Better analysis of bias/variance.

Avoidable bias	Humans	aim here → 1%	7.5%	no room for improvement
	Training error	8%	8%	
variance	Dev error	10%	10%	aim here (reduce)
		focus on bias	focus on variance	

Human level error as the proxy for bayes error.

Understanding human level Performance:

As a proxy for bayes error

↳ least of humans

Surpassing Human level performance:

Team of Humans	0.5 %	0.5 %
Human	1 %	1 %
Training error	0.6 %	0.3 %
Dev error :	0.8 %	0.4 %

↓

⟶ Here you don't know what to optimise, bias or variance.

→ Once you surpass human level, scope for improvement is not clear.

In structured data Machine >> Human.

They are not natural perception problems.

Improving Your model Performance:

2 fundamental Assumption:

1. You can fit training set pretty well.
2. The training set performance generalises pretty well to the dev/test set.

Human level
↕ avoidable bias
Training error
↕ variance
Dev error

- Train bigger model
- Train longer/better opti. algos. – momentum, RMS prop.. Adam
  - ↓ better NN architecture – hyper p. search → RNN, CNN

More data (generalized)
Regularization
(L_2, drop, data augmentation)
NN archi., hyperpara. search