Use ONE real number for metrics

For multiple metrics:
- 1 optimizing (the one to min or maximize).
- N-1 satisficing (accept in range) $\leq, \geq \ldots$

optimizing ↓ ↓ satisficing

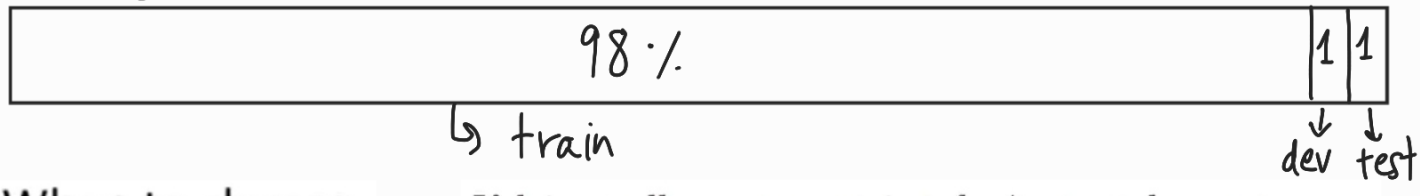
Classifier	Accuracy	Running time
A	90%	80ms
B	92%	95ms
C	95%	1,500ms

Create dev/test sets randomly from ALL the data (so that they come from the same distribution).

Size of test set

→ Set your test set to be big enough to give high confidence in the overall performance of your system.

In big data:



When to change dev/test sets and metrics

→ If doing well on your metric + dev/test set does not correspond to doing well on your application, change your metric and/or dev/test set.

Human-level error (proxy for Bayes error)

"Avoidable bias"
- Train bigger model
- Train longer/better optimization algorithms – momentum, RMSprop, Adam
- NN architecture/hyperparameters search – RNN, CNN

Training error

"Variance"
- More data
- Regularization – L_2, dropout, data augmentation
- NN architecture/hyperparameters search

Dev error

Problems where ML significantly surpasses human-level performance

- Online advertising
- Product recommendations
- Logistics (predicting transit time)
- Loan approvals