Goal of model evaluation:

Determine how well a model building procedure will work for our data.

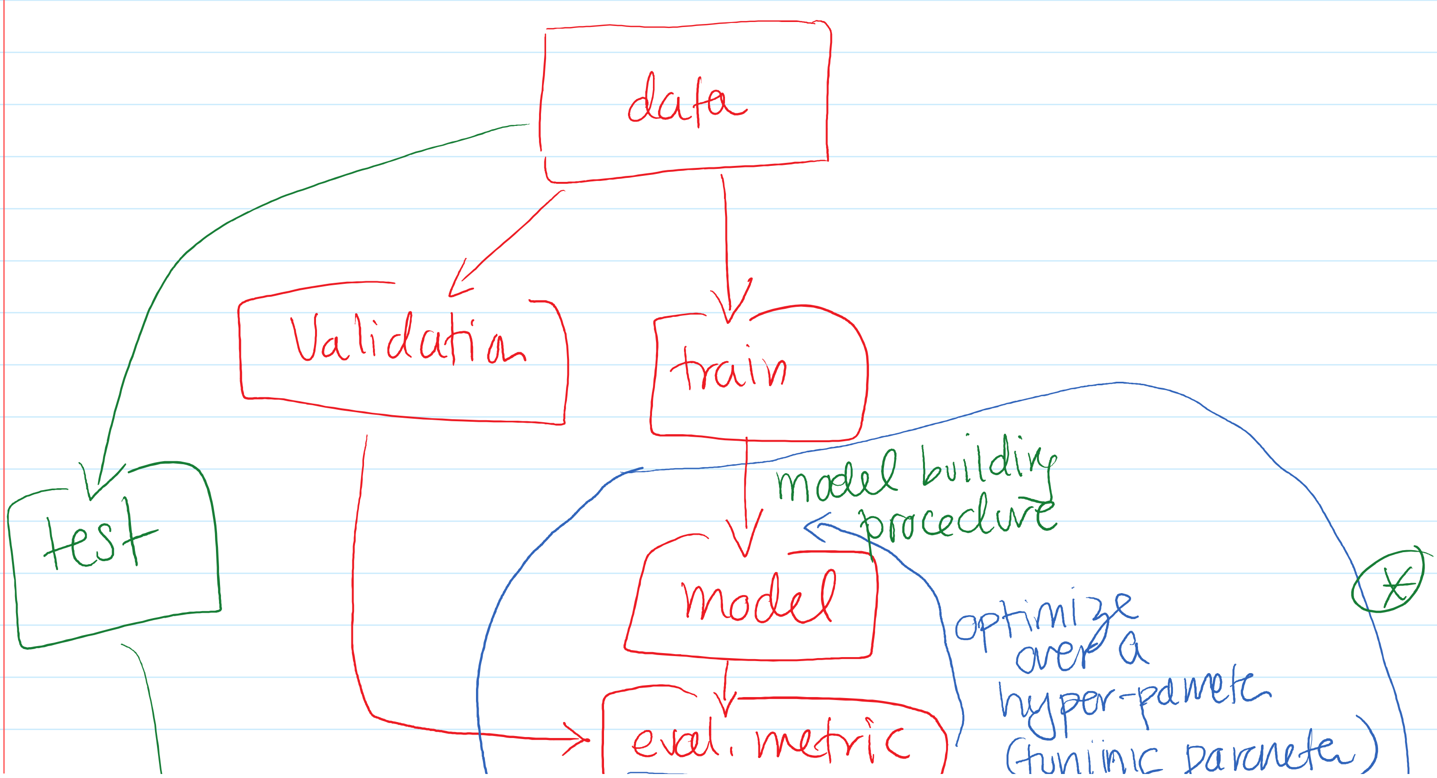
In mind: data follows some prob. dist.

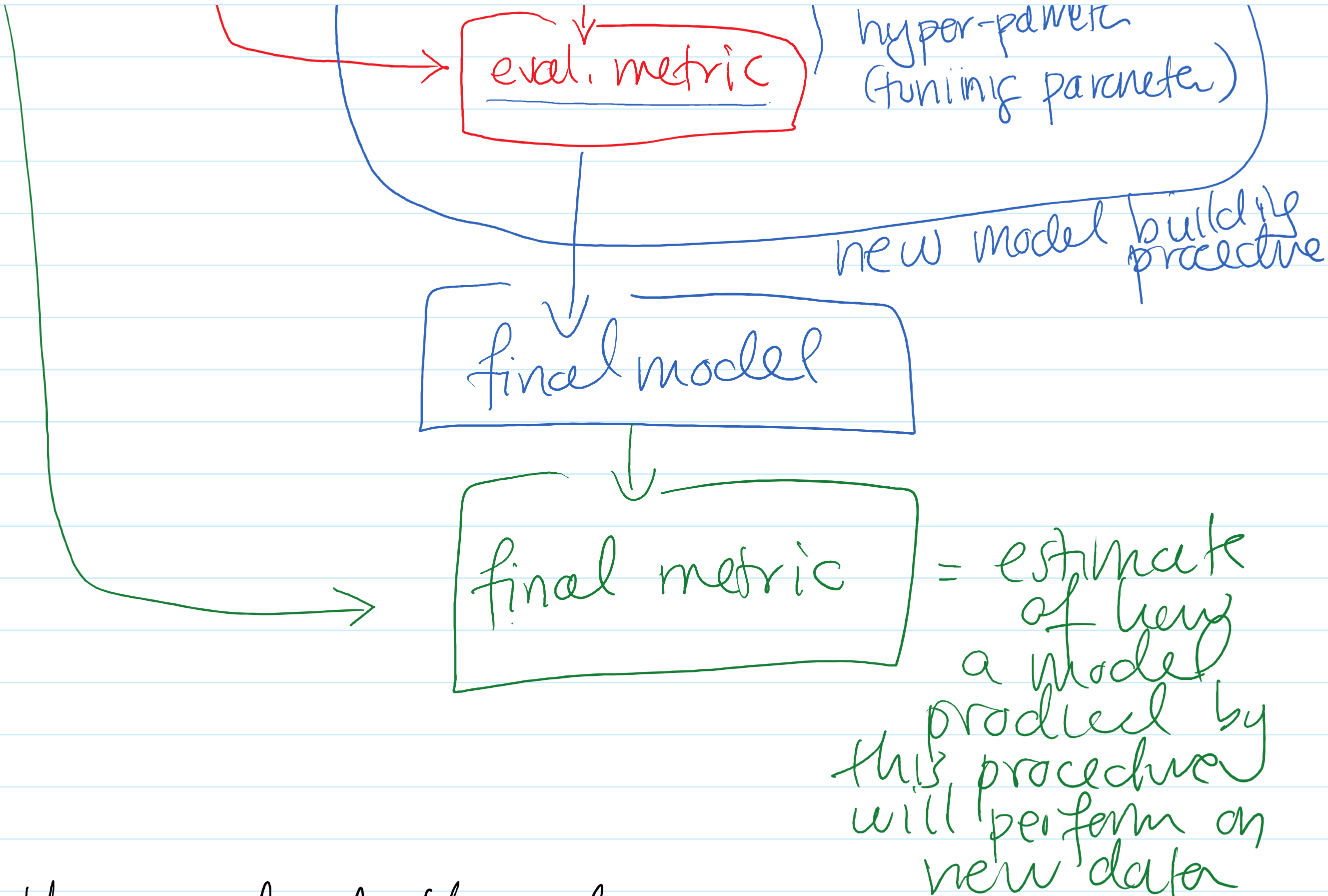
$$\mathbb{D} \in \mathbb{P}$$

data — prob.

→ to form an estimate of how well a model created by a certain procedure will work on unforseen data.

Train/Validation/test





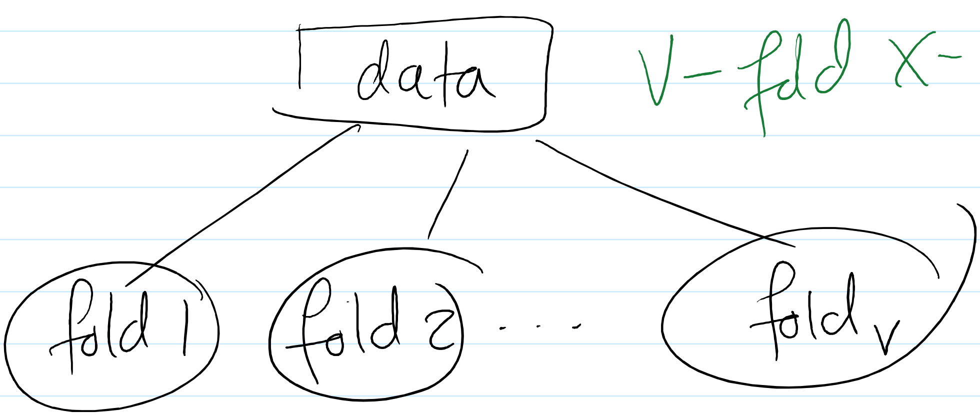
At the end of the day:

you'll build model using all data

X-validation: this has the same problems.

↳ just a fancier way of estimating performance

V-fold X-validation



procedure: For $v = 1, \dots, V$

procedure: For $v = 1, \dots, V$
fit on all but fold k
$m_k \leftarrow$ evaluate on fold k

final metric = mean(m_ks)

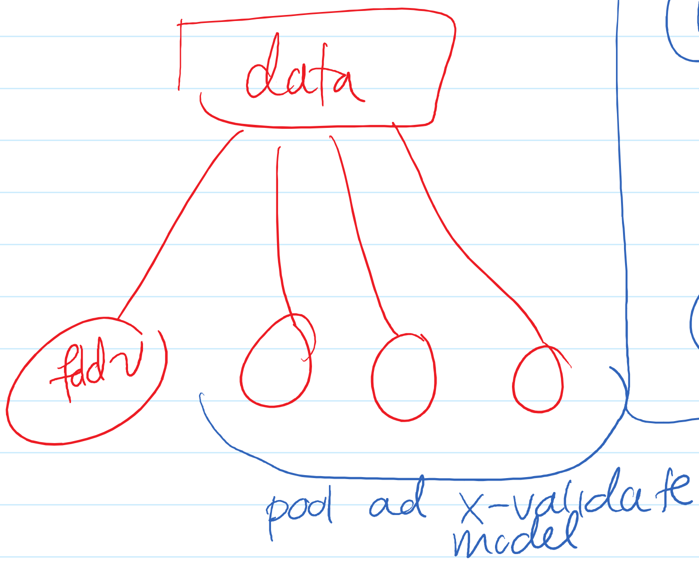
To optimize over a hyper parameter

Nested X-validation:

procedure:

outer loop

For $v = 1, \dots, V$



(1) fit and evaluate using X-validation on all but fold v and optimize over hyper param → final model

inner loop

(2) evaluate on fold k → m_k

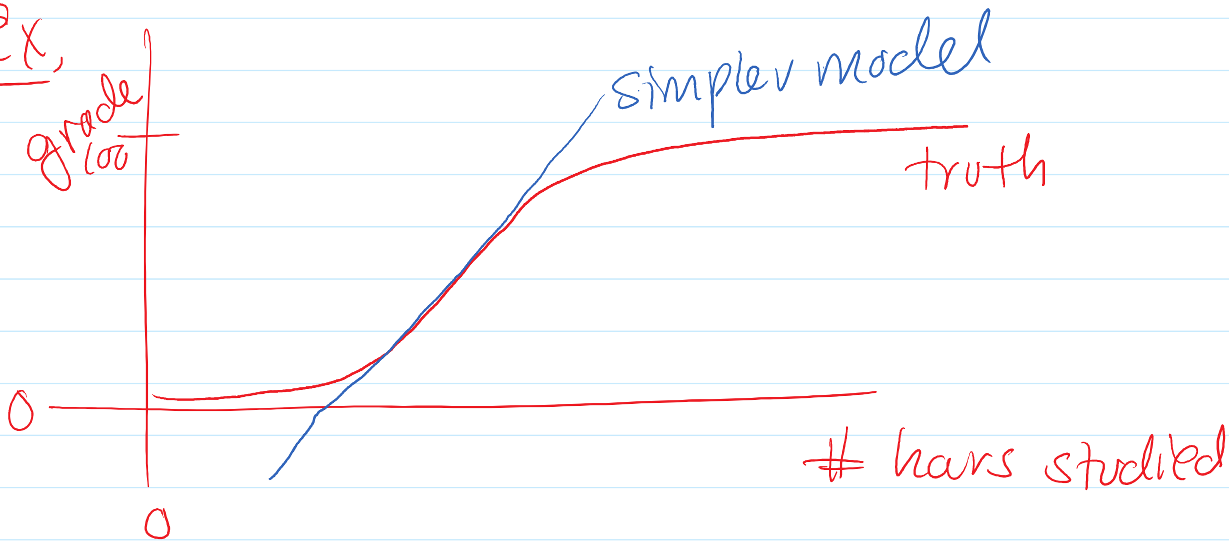
$\boxed{\text{final metric} = \text{avg}(m_k\text{s})}$

Bias/Variance Tradeoff

Laymens Terms:

Bias: error incurs b/c we approx. a complicate real life phenomena w/ a simpler model.

Ex,

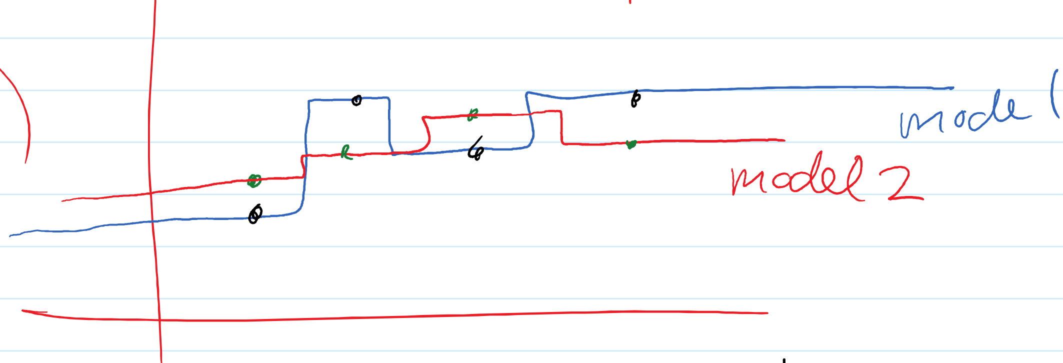


Variance: how much the model fit changes b/c the train data changes.

Ex,

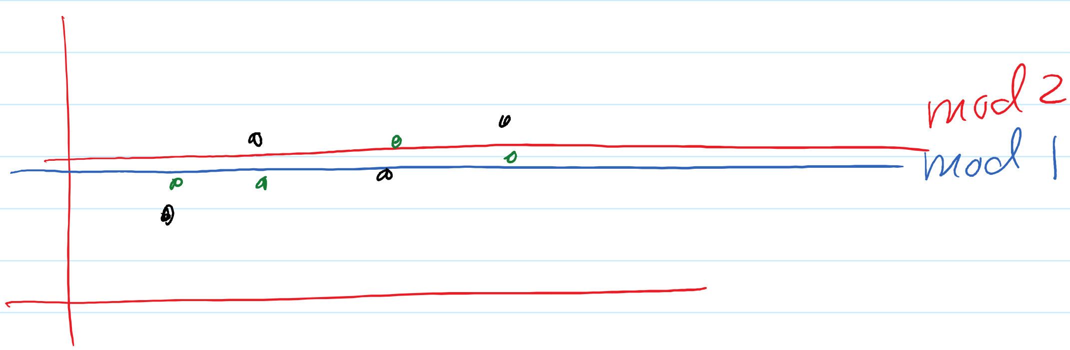
KNN w/ K=1

(high variance)



KNN w/ K=N

(low variance)

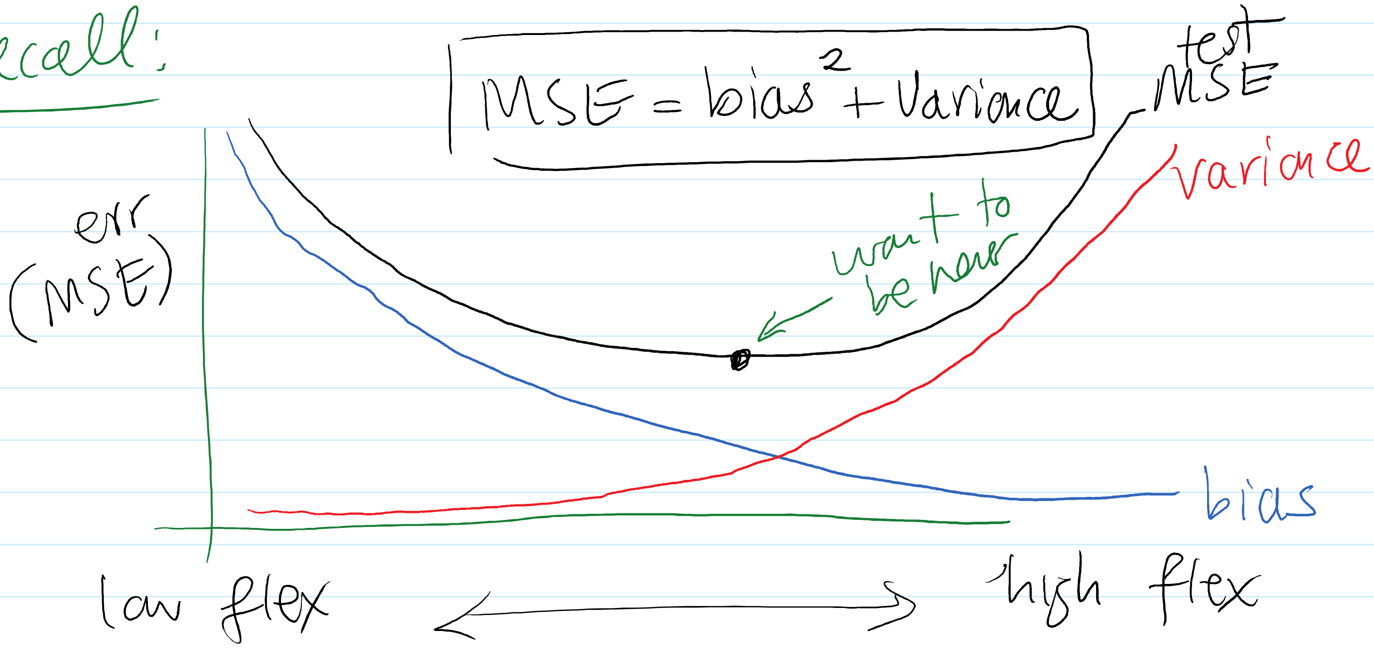


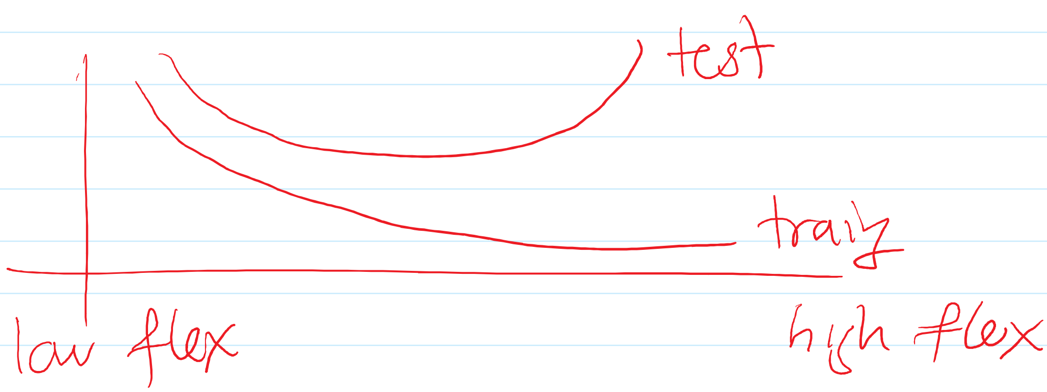
Generally:

tension:

- low variance $\iff$ low flexibility and higher bias
- high variance $\iff$ higher flex, and lower bias

Recall:





$MSE = bias^2 + Variance$

Our data assumed to come from a prob. dist.

dist.

Sample: $MSE = \frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{y}_n)^2$ [sample MSE]

$\rightarrow \quad MSE = \mathbb{E}\left[(y-\hat{y})^2\right]$ [populuta MSE]

random $\qquad\qquad\qquad \hat{y} = \hat{f}(x)$

$\text{Bias}(\hat{f}) = \mathbb{E}[\hat{f}] - f$

expected $\hat{f}$ less truth

$\text{Var}(\hat{f}) \leftarrow$ variability of $\hat{f}$

$y = f(x)$ and $\hat{y} = \hat{f}(x)$

$$MSE = \mathbb{E}\left[(f-\hat{f})^2\right]$$

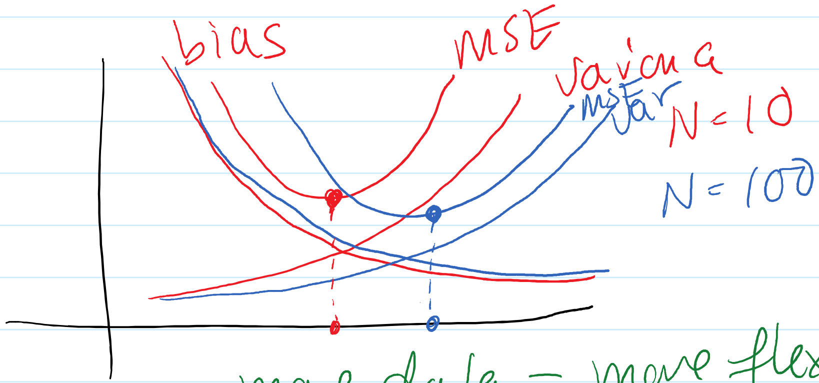
$$= \mathbb{E}\left[\left((f - \mathbb{E}[\hat{f}]) + (\mathbb{E}[\hat{f}] - \hat{f})\right)^2\right]$$

$$= \mathbb{E}\Big[(\underbrace{f - \mathbb{E}[\hat{f}]}_{\text{not random}})^2 + (\mathbb{E}[\hat{f}] - \hat{f})^2 + 2\underbrace{(f - \mathbb{E}\hat{f})}_{\text{not random}}(\mathbb{E}\hat{f} - \hat{f})\Big]$$

$$= \underbrace{\boxed{(f - \mathbb{E}[\hat{f}])^2}}_{\text{Bias}^2} + \underbrace{\boxed{\mathbb{E}\left[(\hat{f} - \mathbb{E}\hat{f})^2\right]}}_{\text{Var}(\hat{f})} + \cancel{2(f - \mathbb{E}\hat{f})}\underbrace{(\mathbb{E}\hat{f} - \mathbb{E}\hat{f})}_{0}$$

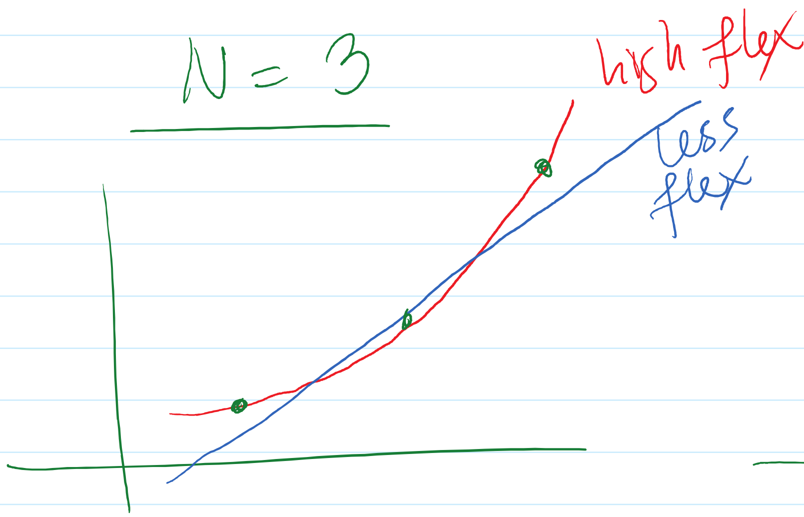
Sample size density:

More data tends to reduce variance.



more data = more flexible reasonable

N = 3



N = 100