SUPERVISED LEARNING

- BIAS-VARIANCE TRADEOFF
- OVERFITTING AND MODEL FLEXIBILITY
- NO FREE LUNCH
- CURSE OF DIMENSIONALITY

EXPECTED PREDICTION ERROR

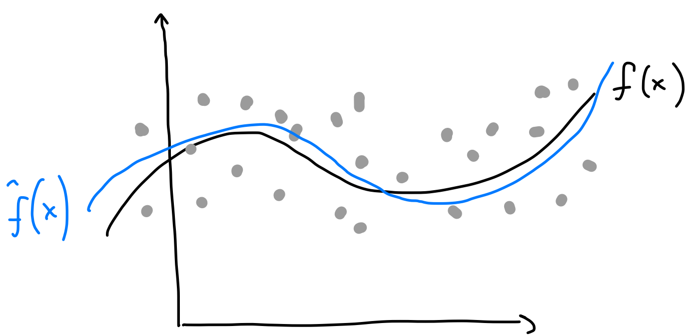
OF PREDICTING $Y = f(X) + \varepsilon$

WITH $\hat{f}(X)$ WHEN $X = x$

MEAN SQUARE ERROR

$\mathbb{V}[\varepsilon]$

$$\mathrm{EPE}\left(Y, \hat{f}(x)\right) = \mathbb{E}_{Y|X,\mathcal{D}}\left[\left(Y - \hat{f}(X)\right)^2 \mid X = x\right] = \underbrace{\mathbb{E}_{\mathcal{D}}\left[\left(f(x) - \hat{f}(x)\right)^2\right]}_{\text{reducible error}} + \underbrace{\mathbb{V}_{Y|X}\left[Y \mid X = x\right]}_{\text{irreducible error}}$$

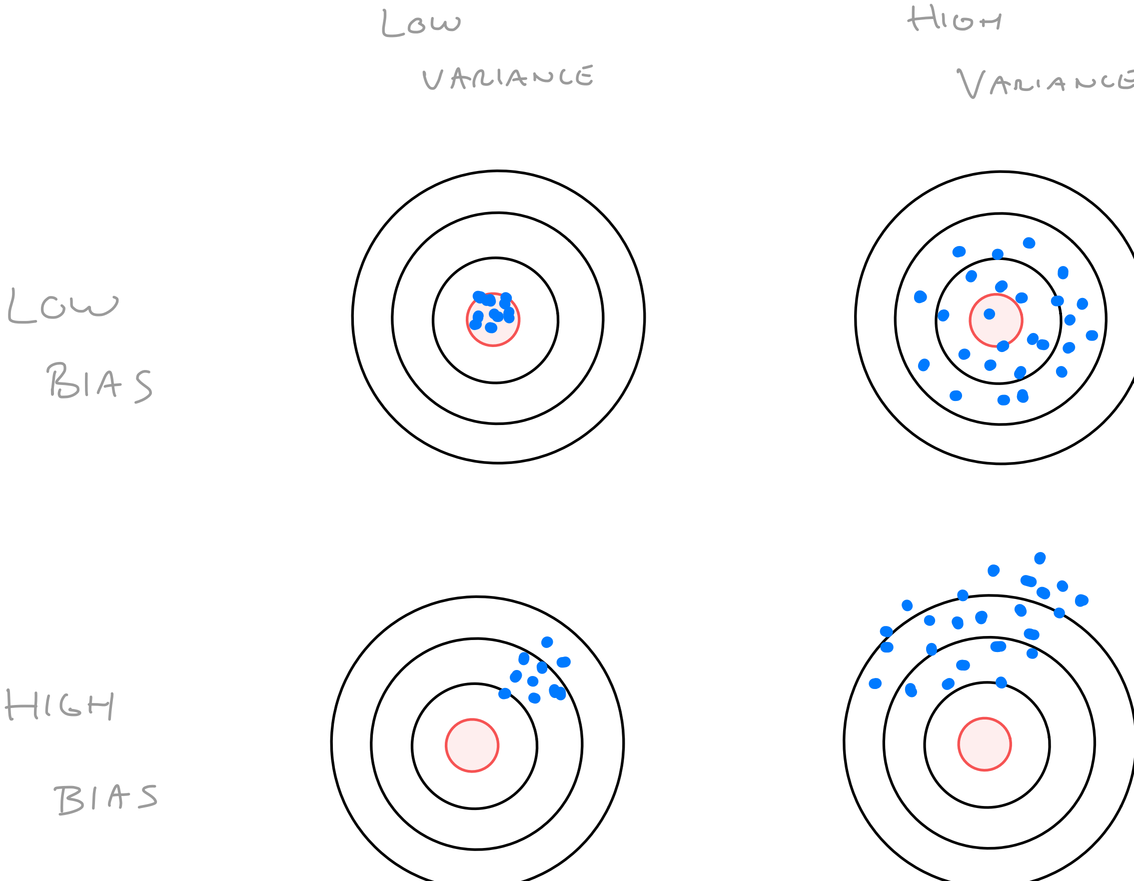


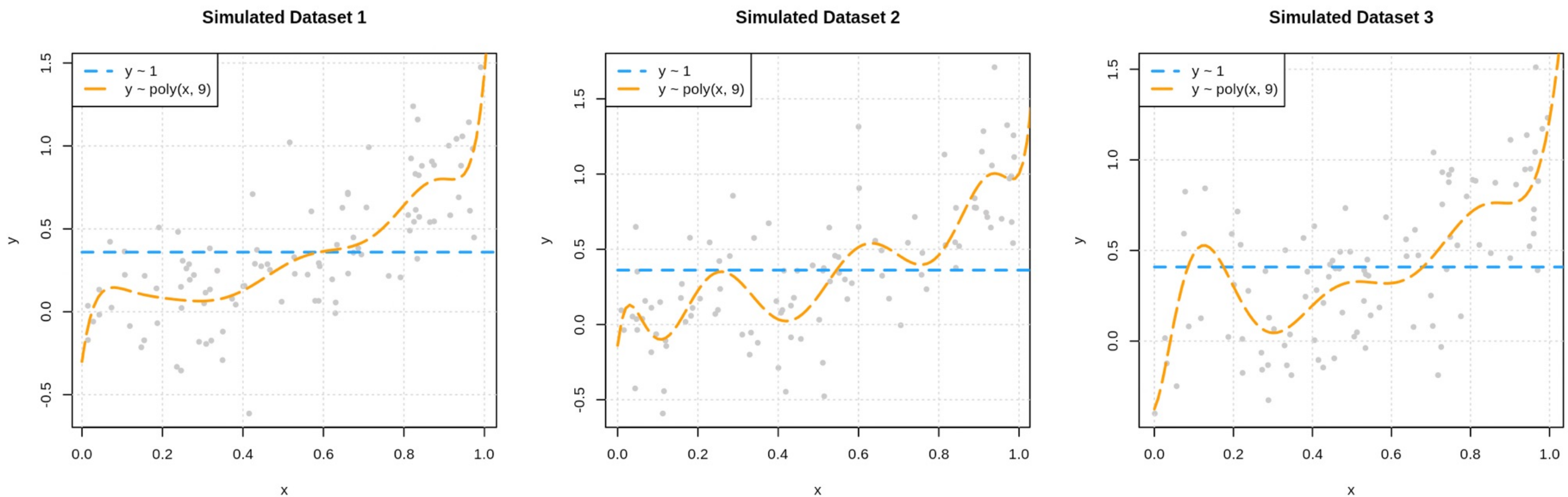
Bias and Variance

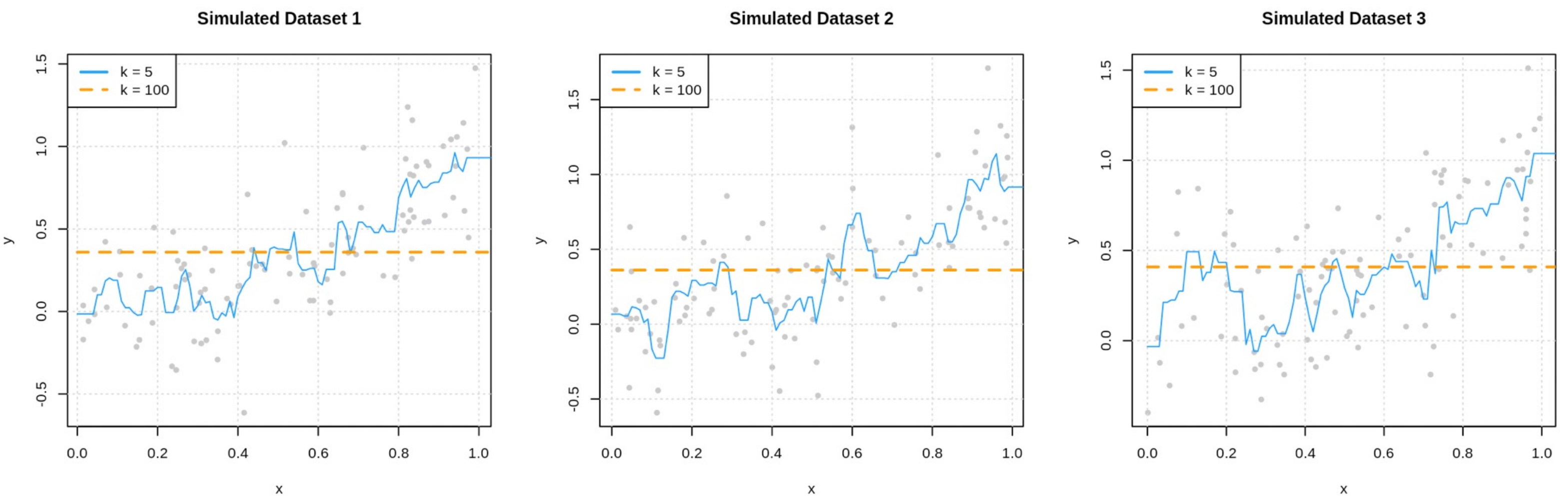
$$\text{Bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$$

$$\text{Var}(\hat{\theta}) = \mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}[\hat{\theta}]\right)^2\right]$$

INSERT JOKE HERE


LOW VARIANCE
HIGH VARIANCE
LOW BIAS
HIGH BIAS


Simulated Dataset 1
y ~ 1
y ~ poly(x, 9)
y
x
Simulated Dataset 2
y ~ 1
y ~ poly(x, 9)
y
x
Simulated Dataset 3
y ~ 1
y ~ poly(x, 9)
y
x



Simulated Dataset 1
Simulated Dataset 2
Simulated Dataset 3
k = 5
k = 100
x
y

MEAN SQUARED ERROR

$$MSE\left(f(x), \hat{f}(x)\right) = BIAS^2\left(\hat{f}(x)\right) + VAR\left(\hat{f}(x)\right)$$

BIAS VARIANCE TRADEOFF

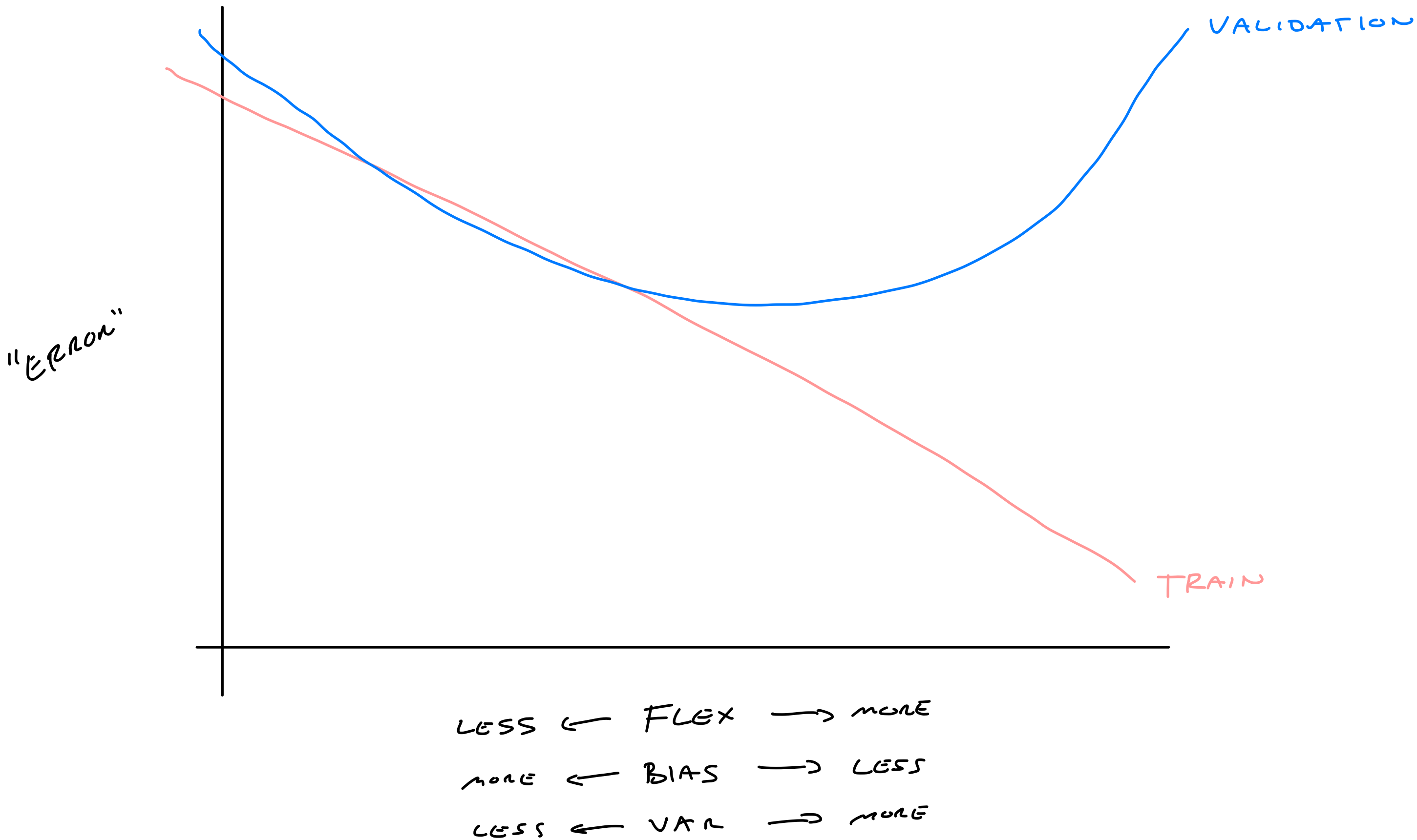
AS BIAS $\uparrow$, VAR $\downarrow$

BUT NOT THE SAME RATE

MODEL FLEXIBILITY

MODELS THAT ARE MORE "WIGGLY" ARE MORE FLEXIBLE.

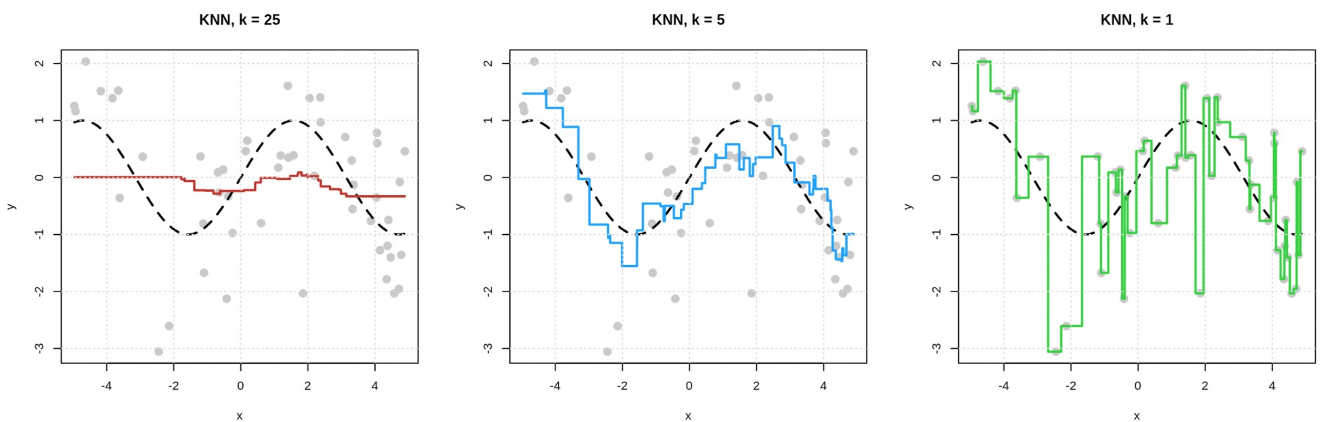
THESE MODELS ARE "VARIABLE."

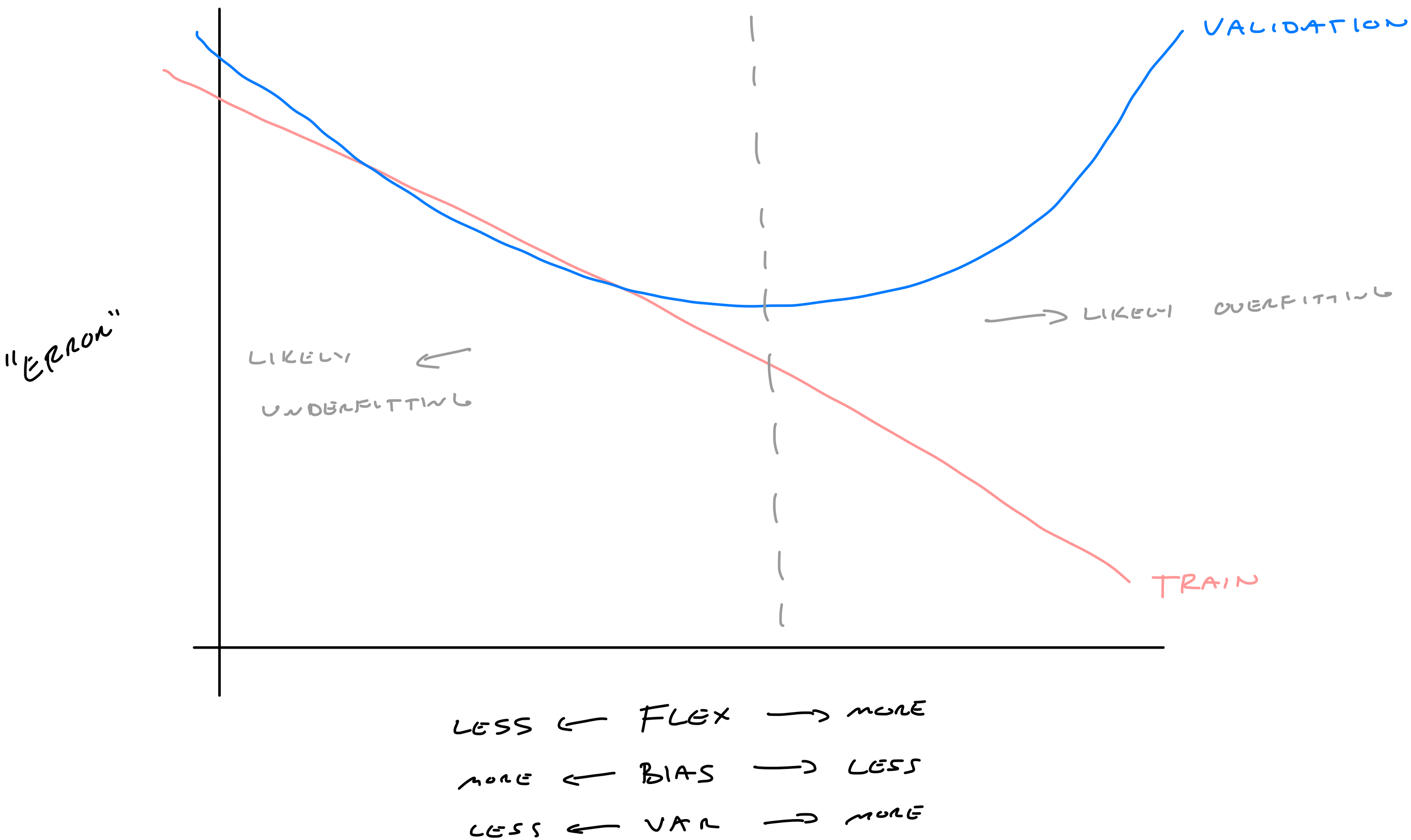

"ERROR"
VALIDATION
TRAIN
LESS ← FLEX → MORE
MORE ← BIAS → LESS
LESS ← VAR → MORE

OVERFITTING

WHEN A MODEL IS TOO FLEXIBLE

- TRAIN ERROR WILL BE OPTIMISTIC
- "FIT TO NOISE"




VALIDATION
LIKELY OVERFITTING
"ERROR"
LIKELY
UNDERFITTING
TRAIN
LESS ← FLEX → MORE
MORE ← BIAS → LESS
LESS ← VAR → MORE

NO METHOD WILL PERFORM BEST ON ALL POSSIBLE DATASETS.

NO FREE LUNCH

Curse of Dimensionality

In high dimensions, any datapoint will have no "close" neighbours.

n = sample size

p = # features