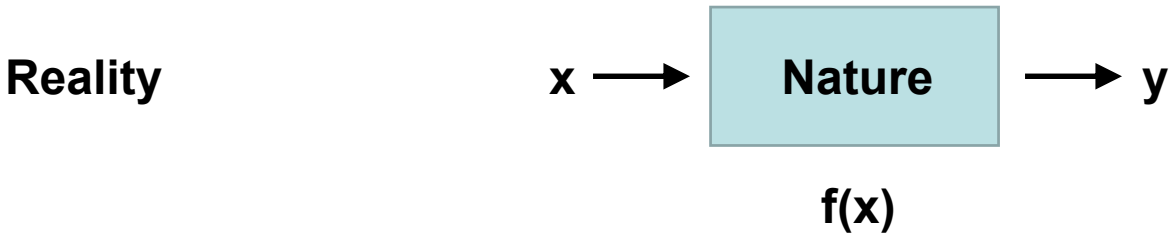
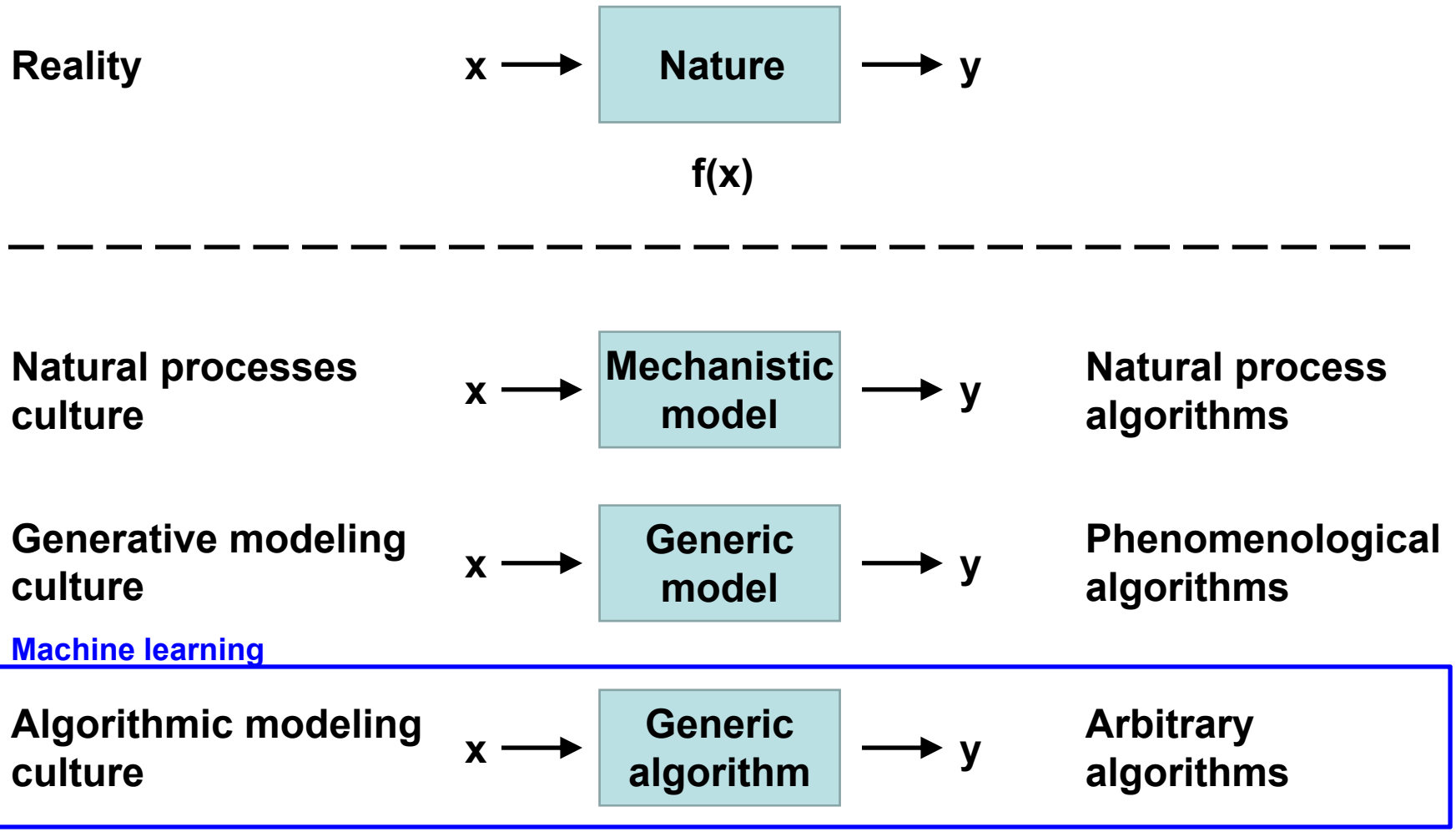
Trying to learn a function f



Trying to learn a function f



f can mean different things in different cultures

f(x) for prediction

Reality $Z_1, Z_2, \ldots, Z_\Omega$ $Y = g(Z)$

Some set of causally-connected variables

We have $X_1, X_2, \ldots, X_p$

A set of potential predictor variables

$$Y = f(X) + \epsilon$$

Systematic component Error

Prediction $\hat{Y} = \hat{f}(X)$ Hats indicate predicted Y and estimated f

Goal of prediction

Use data to find a function $\hat{f}$ that has good predictive performance given X

That is, $\hat{f}$ is accurate on new observations

Goal of machine learning

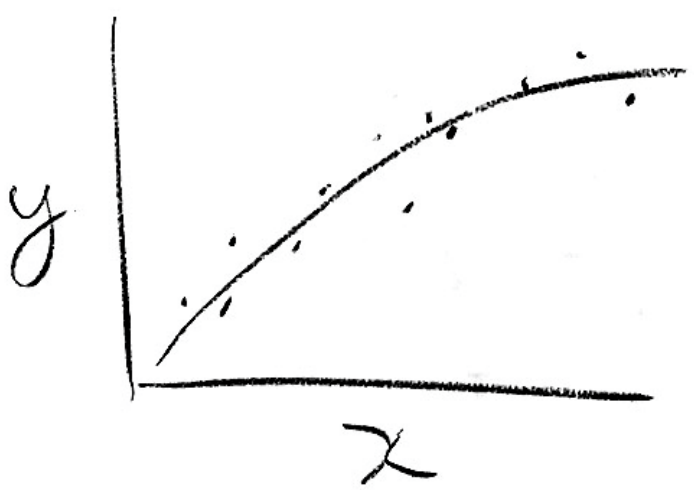
To predict accurately!

- Species distribution
  - map
  - predict accurately for places we won't visit
- Climate change forecast
  - predict accurately for the future
- Antelopes in camera trap images
  - hand over the identification task to a machine so we don't have to look at images!
  - predict accurately for images that we'll never look at

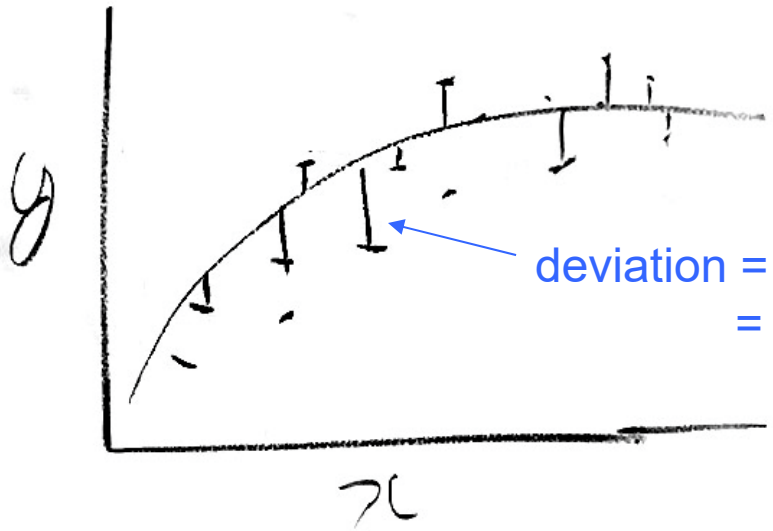
Predictive performance

Basic idea: out-of-sample accuracy

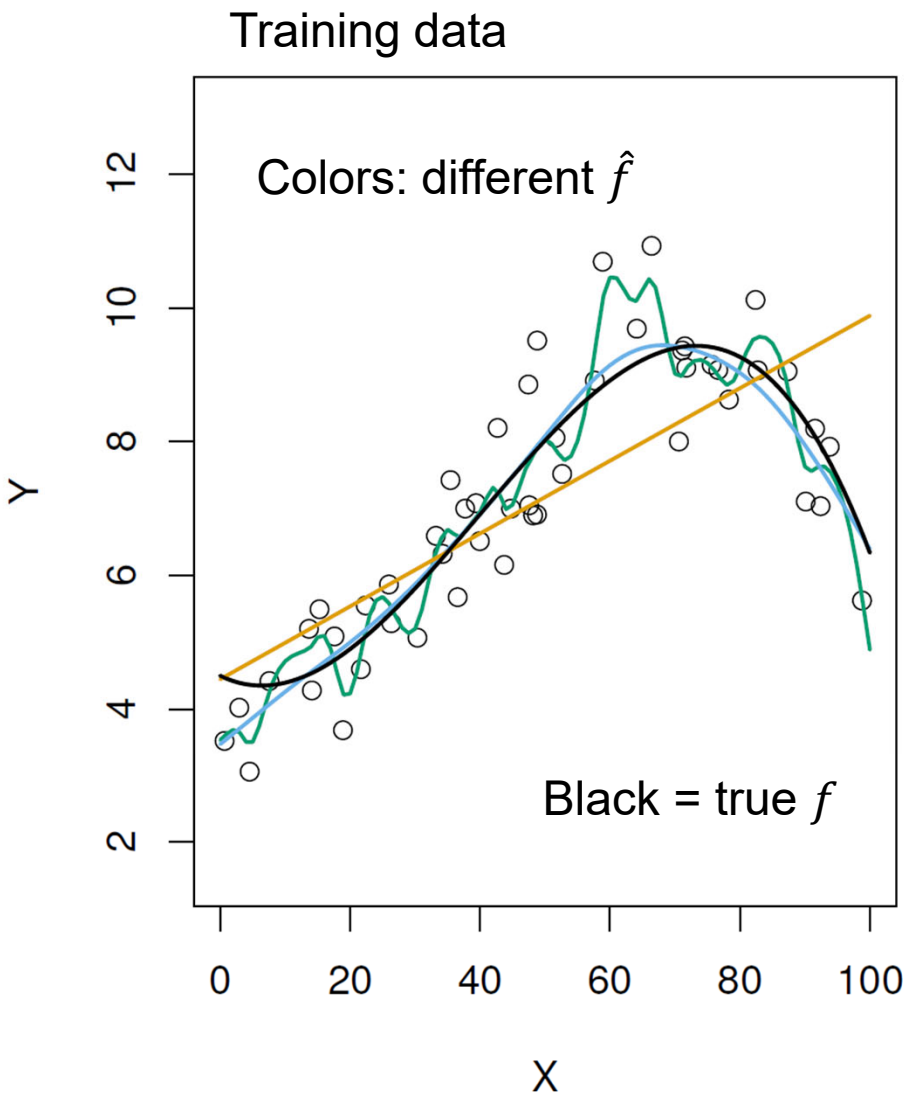
$\hat{f}$ fitted on training data

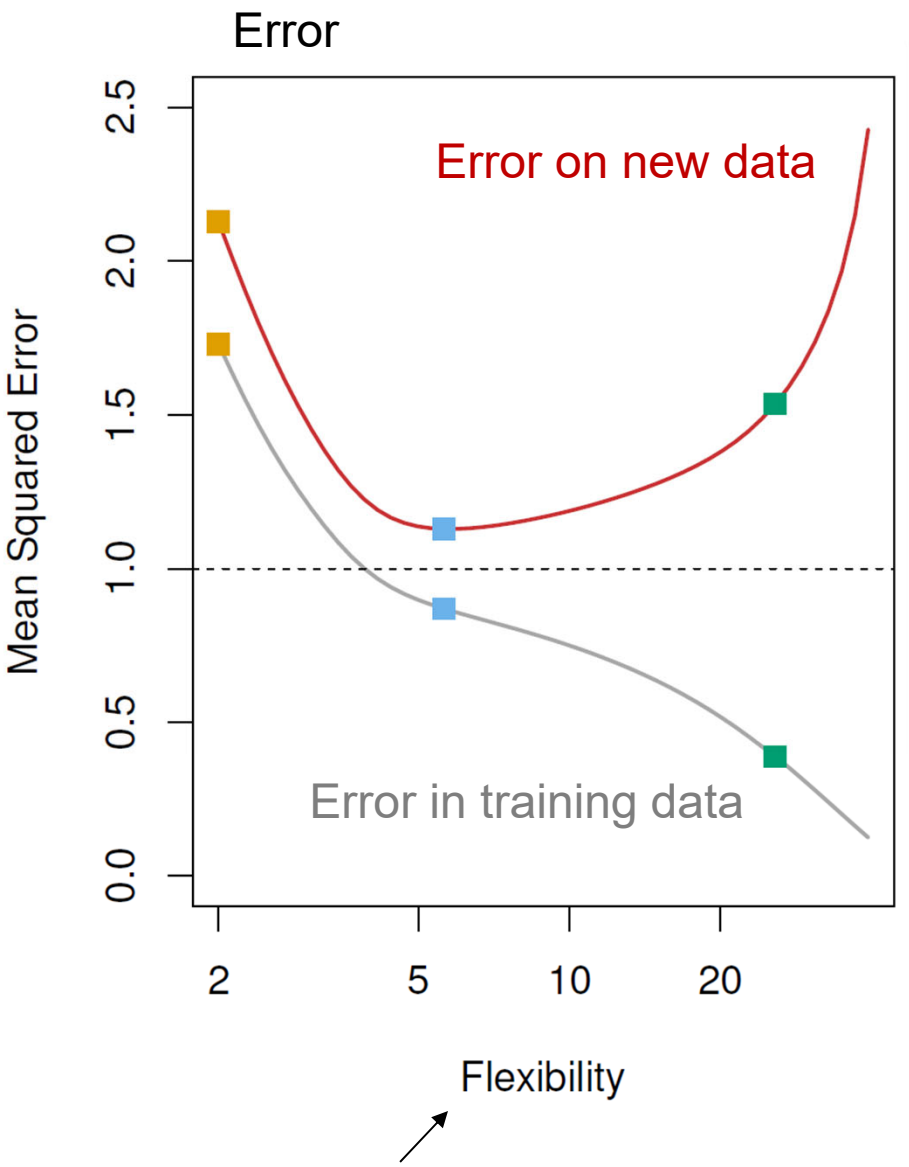


$\hat{f}$ predicting new data

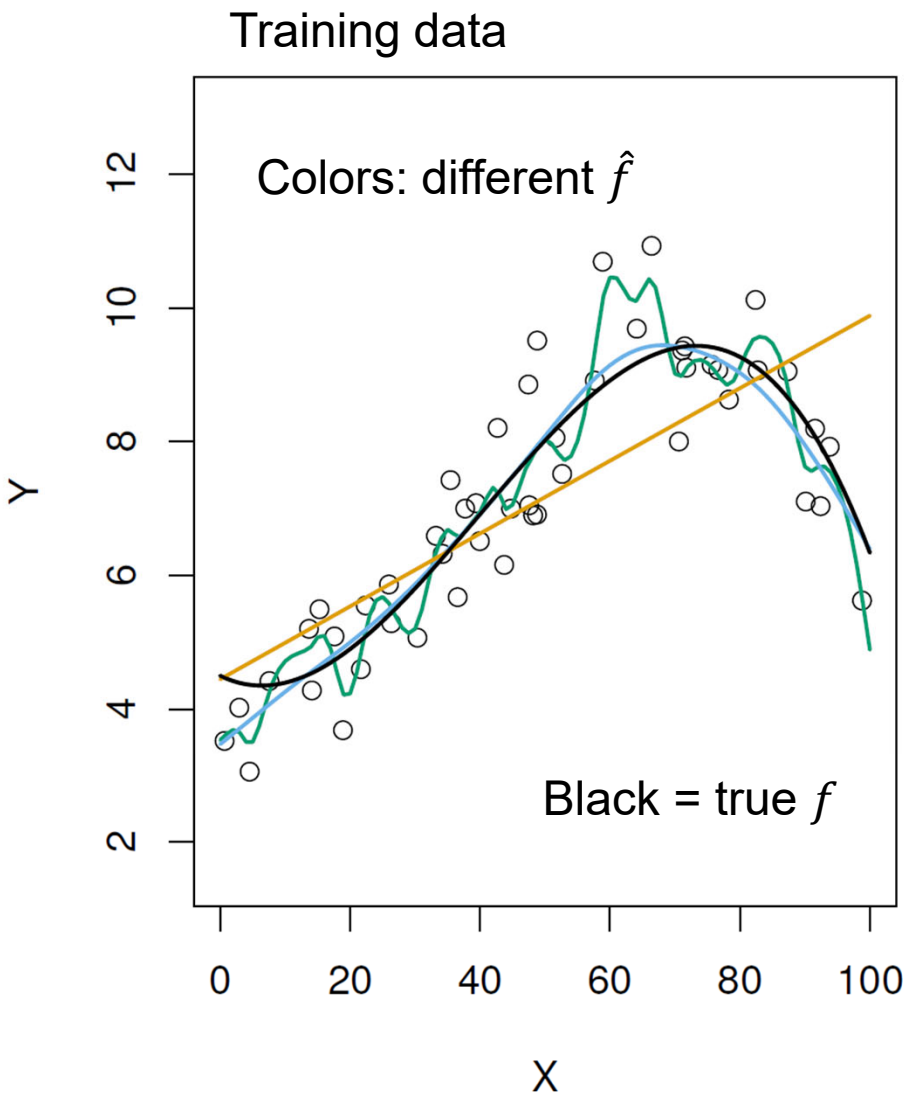


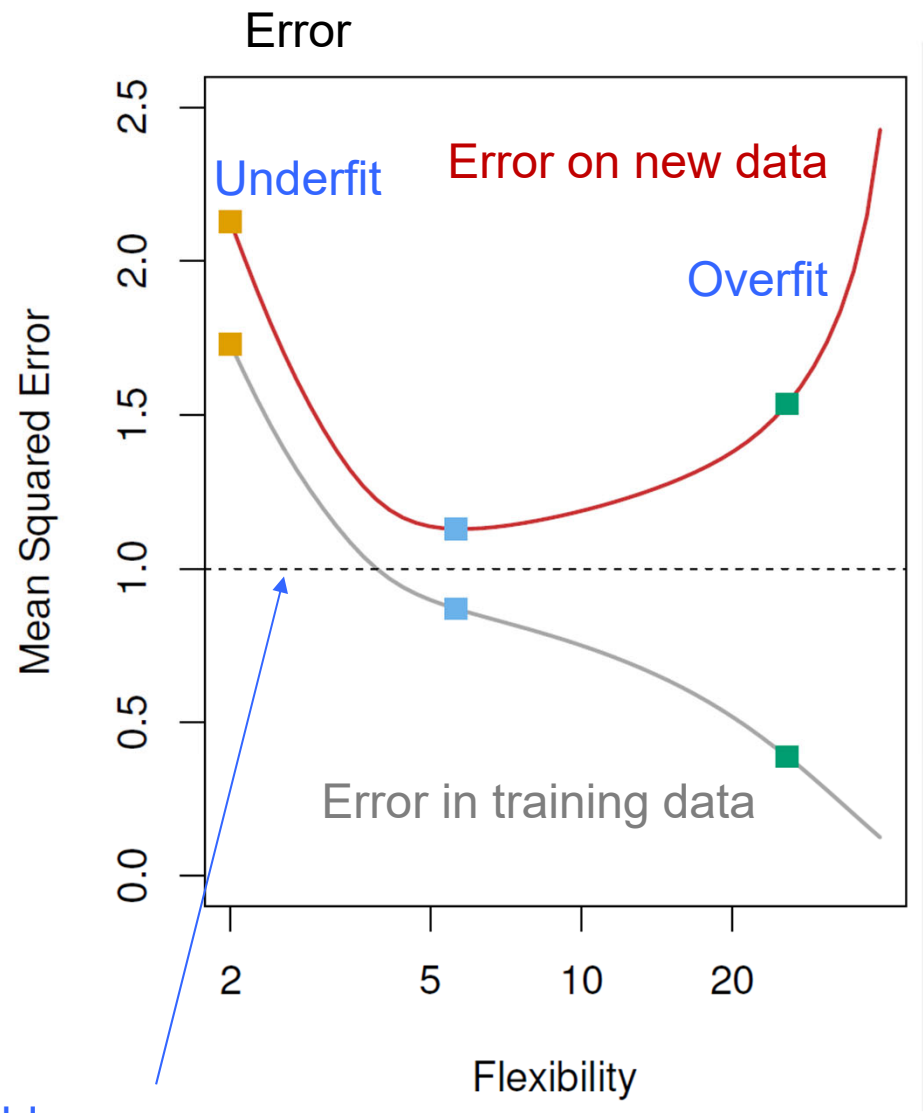
e.g. mean square error (MSE) $\frac{1}{n}\sum_{i}^{n} e_i^2$





Effective d.f. = amount of wiggliness in $\hat{f}$





Goal: balance underfit and overfit

Bias-variance tradeoff

$$\text{Total error} = \text{Var}\left(\hat{f}\right) + \text{Bias}\left(\hat{f}\right)^2 + \text{Var}(\epsilon)$$

Amount $\hat{f}$ would change if fitted on different data

Average deviation of $\hat{f}$ from true f

Irreducible error

Bias-variance tradeoff

$$\text{Total error} = \text{Var}\left(\hat{f}\right) + \text{Bias}\left(\hat{f}\right)^2 + \text{Var}(\epsilon)$$

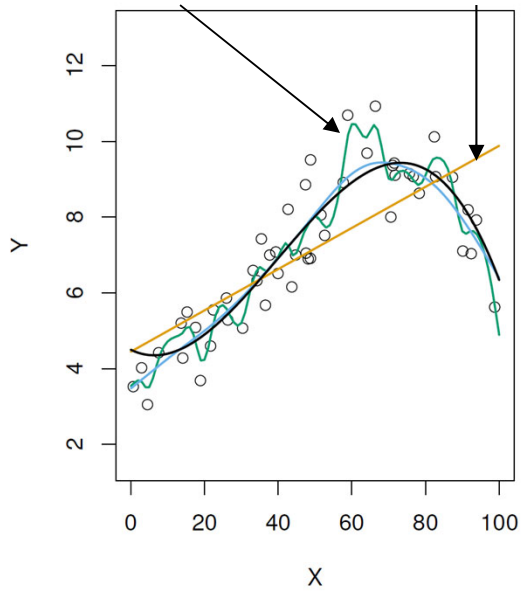
Amount $\hat{f}$ would change if fitted on different data

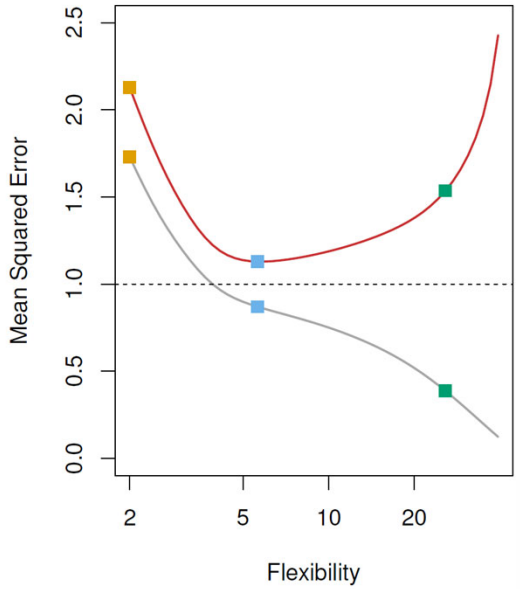
Average deviation of $\hat{f}$ from true f

Irreducible error

High variance

High bias





Bias-variance tradeoff

$$\text{Total error} = \text{Var}(\hat{f}) + \text{Bias}(\hat{f})^2 + \text{Var}(\epsilon)$$

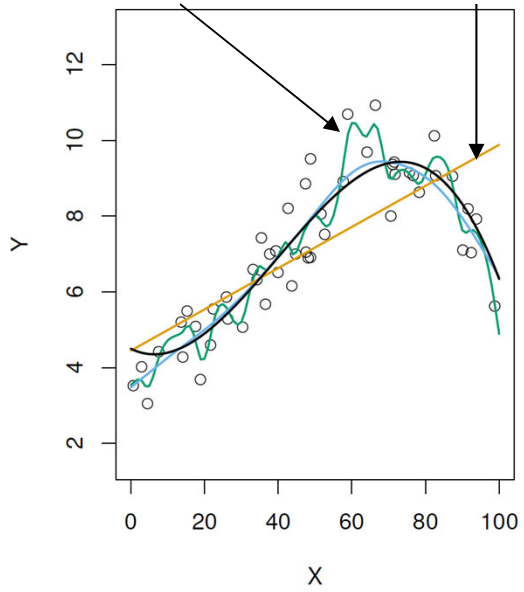
Amount $\hat{f}$ would change if fitted on different data

Average deviation of $\hat{f}$ from true f

Irreducible error

High variance

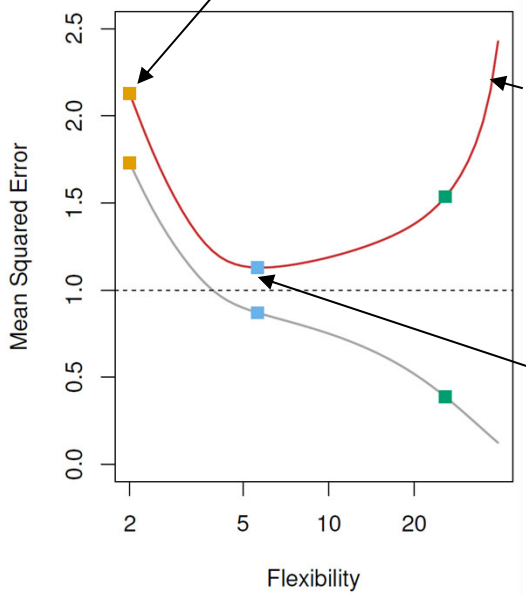
High bias



High bias

High variance

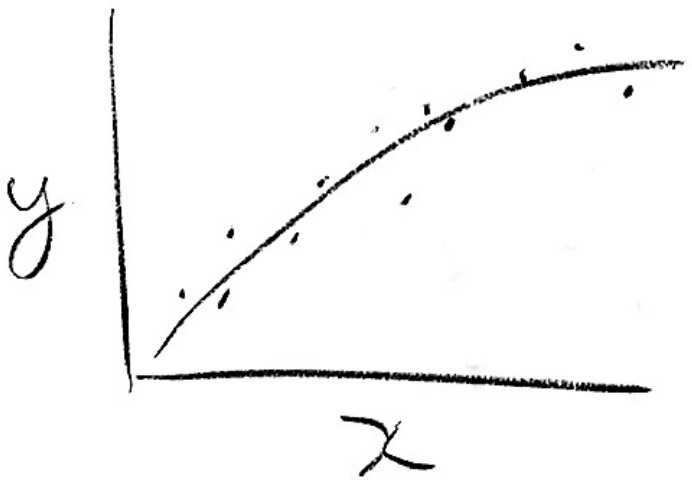
Sweet spot:
low variance
low bias



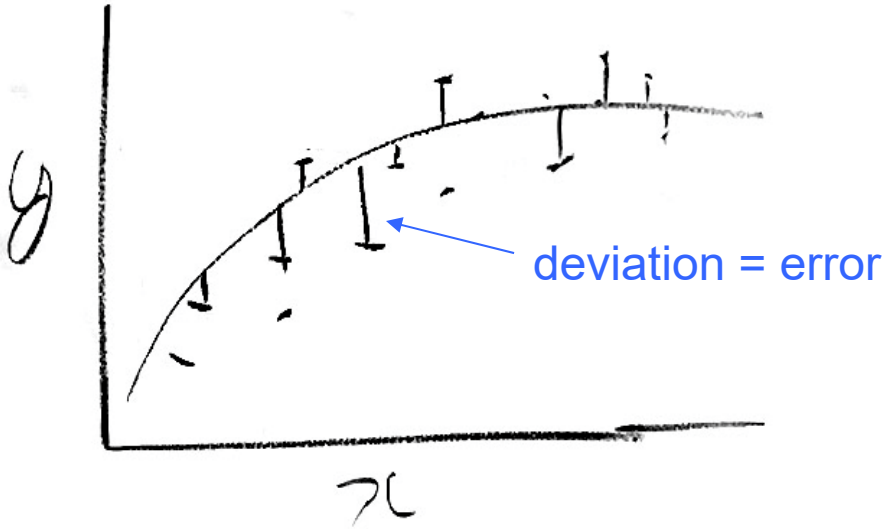
Inference algorithm

Basic idea: out-of-sample validation

Fit model to training dataset



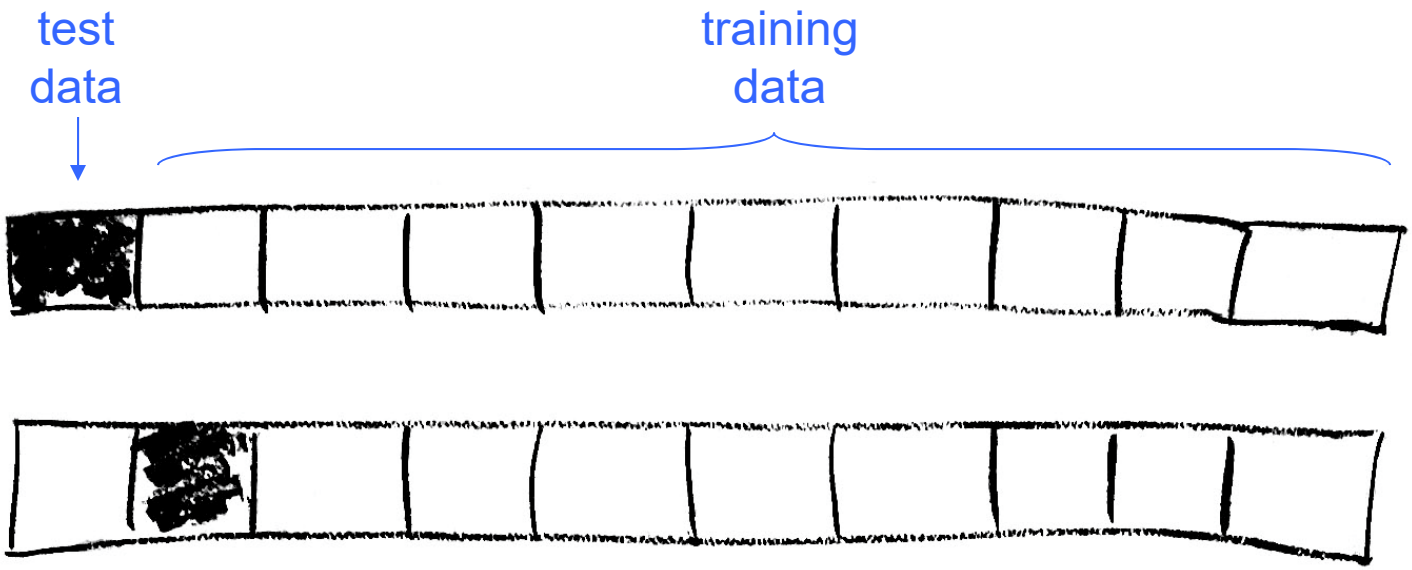
Test model on validation dataset



e.g. mean square error (MSE)

k-fold cross validation (CV)

Divide dataset into k parts (preferably randomly)



... repeat with each test subset

k-fold CV inference algorithm

Algorithm

divide dataset into k parts i = 1...k
for each i
    test dataset = part i
    training dataset = remaining data
    find f using training dataset
    use f to predict for test dataset
    e_i = prediction error
CV_error = mean(e)

Typical values for k: 5, 10, k

Leave-one-out cross validation

- LOOCV
- = k-fold CV for k = n

Algorithm

for each data point
    fit model without point
    predict for that point
    measure prediction error (compare to observed)
CV_error = mean error across points

(AIC and Bayesian WAIC are equivalent asymptotically)