Cross Validation and Bias Variance

Goals

- Be able to…
  - explain how to get an honest sense of how well a model is doing
    - explain train/test splits
    - state the purposes of Cross Validation
    - explain k in “k-fold cross validation”
  - describe the sources of model error
    - describe overfitting and underfitting
  - explain how to compare models and select the best one

We want to **<u>Do The Best</u>**™

- How can we *evaluate* how well we're doing?
  - Ideally with real numbers
  - Ideally in a comparable way
- How can we *change* how well we're doing?
  - Ideally in a directed, intentional way

How can we evaluate how well we’re doing?

- Story:
    - Timmy learns multiplication

How can we evaluate how well we're doing?

- Memorization vs Generalization
- We usually care about generalization
  - (Otherwise use a database)
  - Eg: predicting stock market data, predicting heart disease
    - we care about 'unseen' data points

How can we evaluate how well we're doing?

- Story:
  - Timmy does well on a test

How can we evaluate how well we’re doing?

- We can set aside some data points to be ‘unseen’ so that, using them, we can evaluate our ability to generalize (rather than memorize)

- Train/Test split

    - How much in each?

        - More in Train: better model, less able to evaluate it well

        - More in Test: worse model, better able to evaluate it well

How can we evaluate how well we're doing?

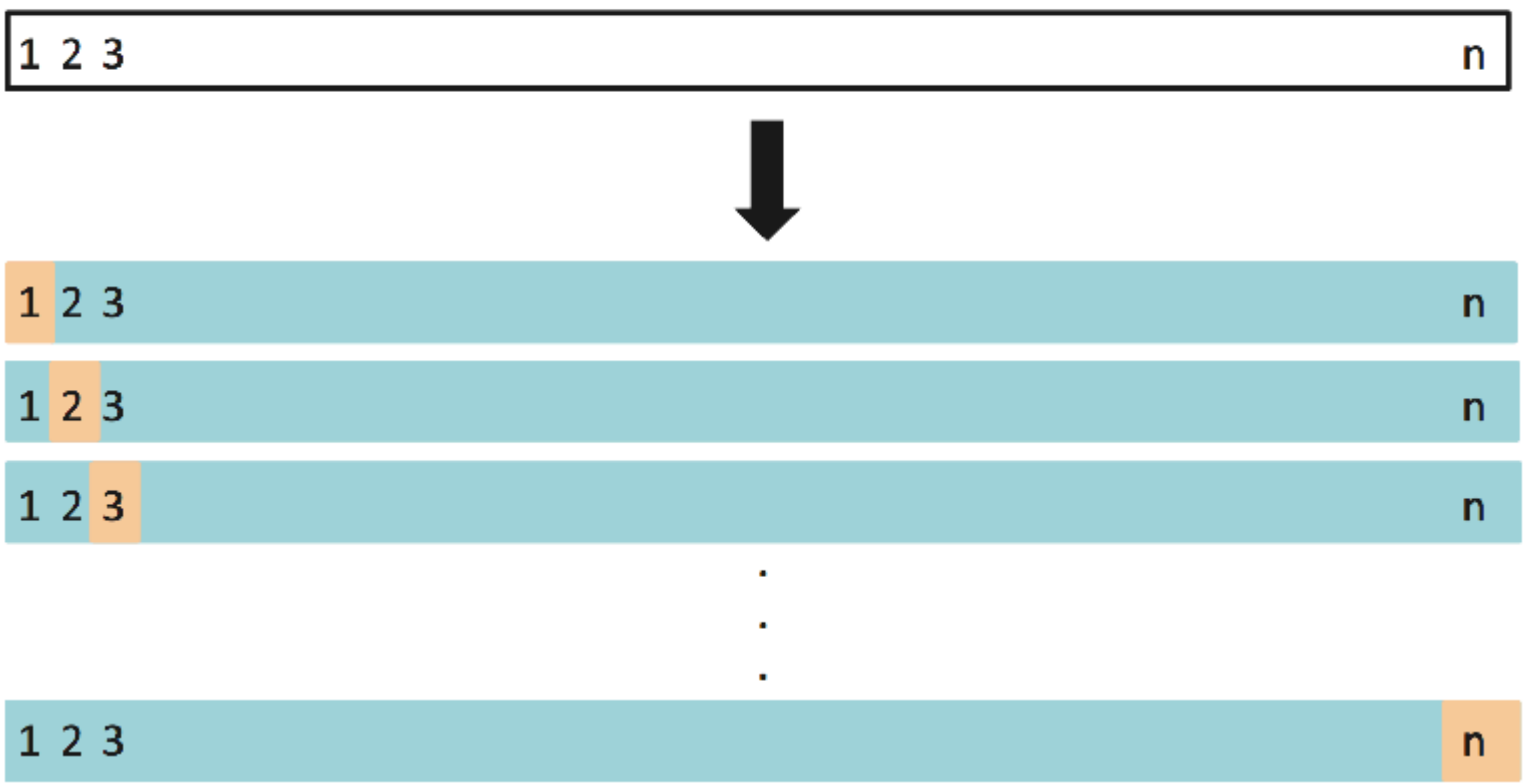
- We could keep the data all together (no, Timmy!)
- We could split the data evenly in two equal-sized groups
- We could split the data into n groups (each data point in its own group)
- We could split the data into k << n groups

How can we evaluate how well we're doing?

- We could split the data evenly in two equal-sized groups

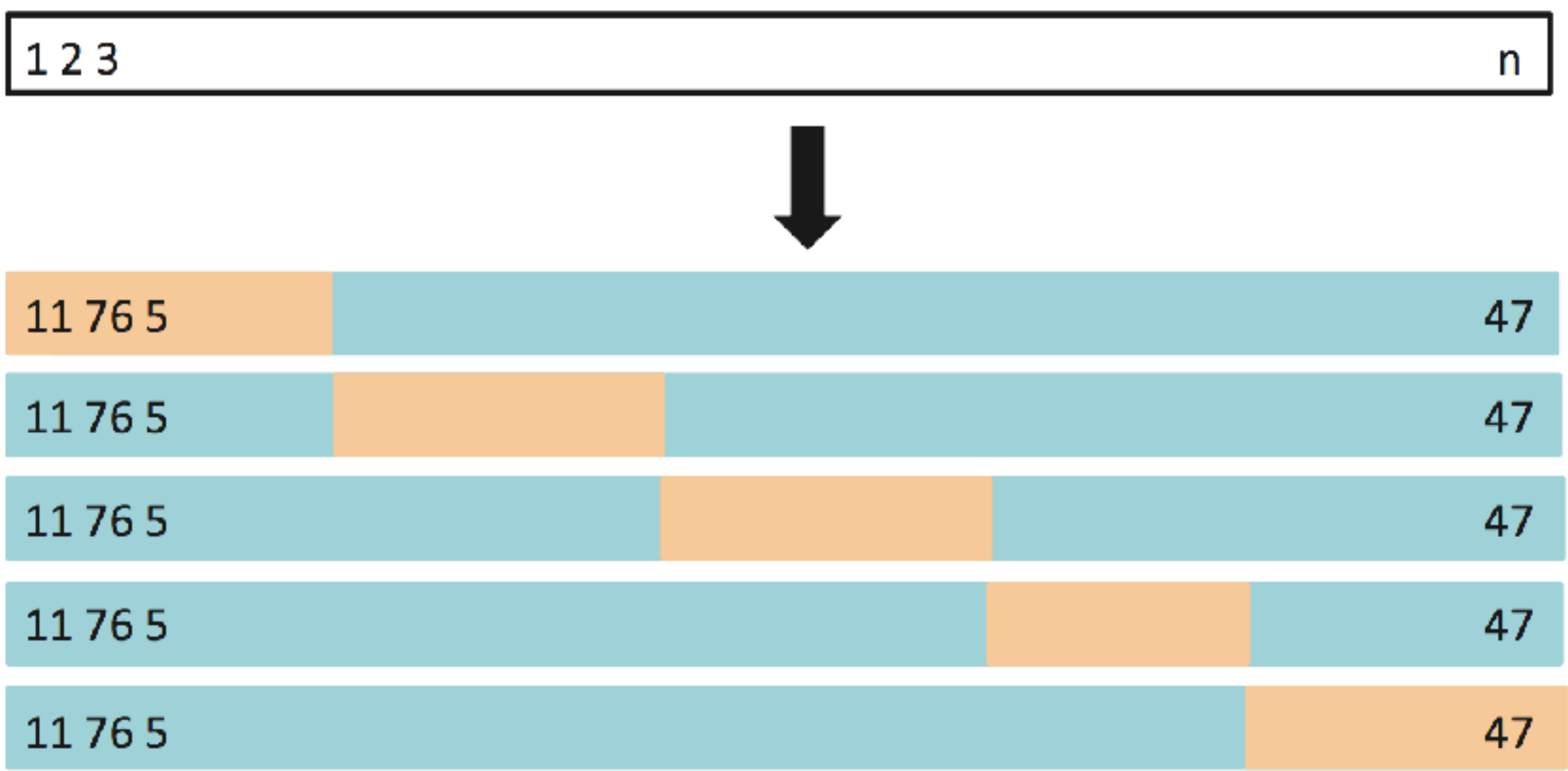
How can we evaluate how well we're doing?

- We could split the data into n groups (each data point in its own group)



How can we evaluate how well we're doing?

- We could split the data into k << n groups



How can we evaluate how well we're doing?

- K Fold Cross Validation:
  - Split dataset D into subsets, D_1, D_2, D_3, … D_k
  - errors_list = []
  - For each subset D_i
    - Mark D_i as the test set
    - Train a model on data in all subsets != D_i
    - errors_list.append(Evaluate model on D_i)
  - Report error = mean(errors_list)

How can we evaluate how well we're doing?

- Story:
  - Timmy can consistently do well on unseen data

How can we change how well we're doing?

How can we change how well we’re doing?

- Try different parameters (broadly speaking)
  - feature presence
    - what do we capture? do we have an ‘is_married’ feature?
  - feature encoding
    - do we encode age as int? float? categorical? what units?
  - feature degree
    - should a feature interact with itself? x_0^2
    - should a feature interact with another? $x_0 * x_{15}$
    - how many features should we have?

How can we change how well we're doing?

- Let's focus on degree

- $y = \beta_0 + \beta_1 x_1$

- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2$

- $y = \beta_0 + \beta_1 x_1 + \beta_1 x_1^2 + \beta_2 x_1^3 + \ldots$

- How do we choose?

How can we change how well we're doing?

- What's the best degree?
- It's not a parameter, exactly, it's a… Hyperparameter!
  - Models usually optimize parameters internally
    - To, say, minimize error on the training set
  - We usually optimize hyperparameters with CV

How can we change how well we're doing?

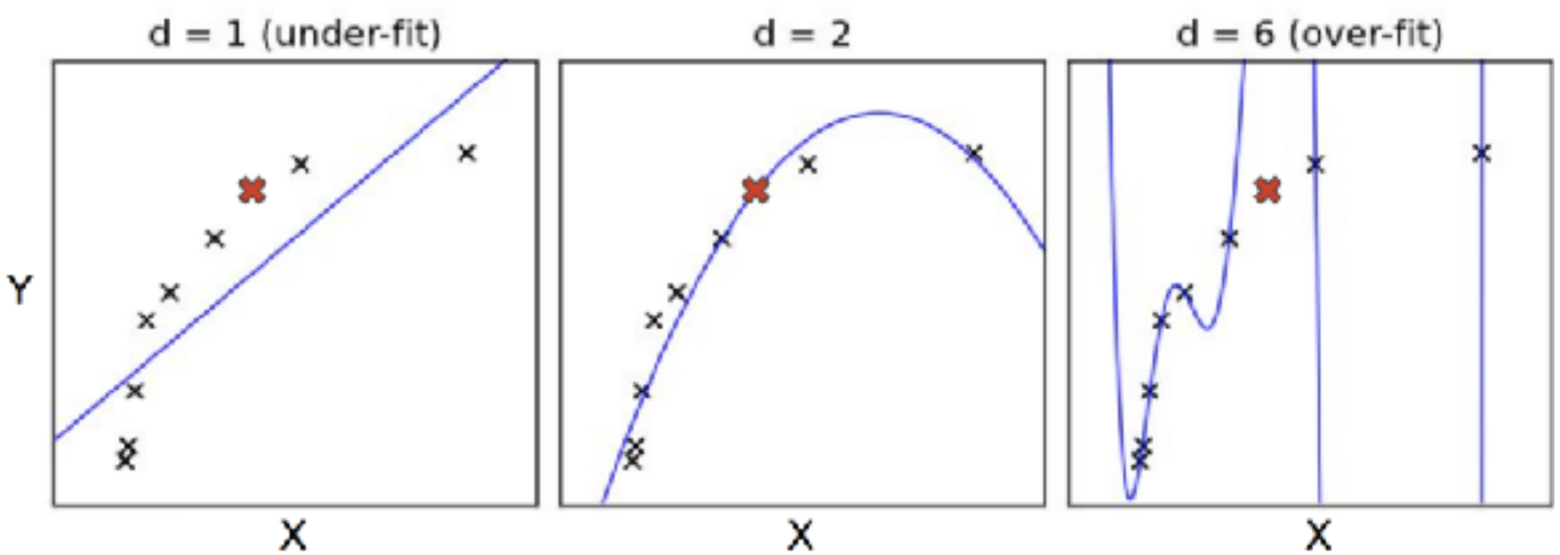
- errors = {}
- For each value *v* of the hyperparameter *h* we want to test…
  - Run a k-fold CV where *h*=*v* (for every model in every fold). Let the mean of the errors be *e*
  - errors[v] = *e*
- Pick the dictionary key *v* w the lowest value *e*
- How well does our model work?
  - How does it perform on unseen data?
  - What data is unseen?

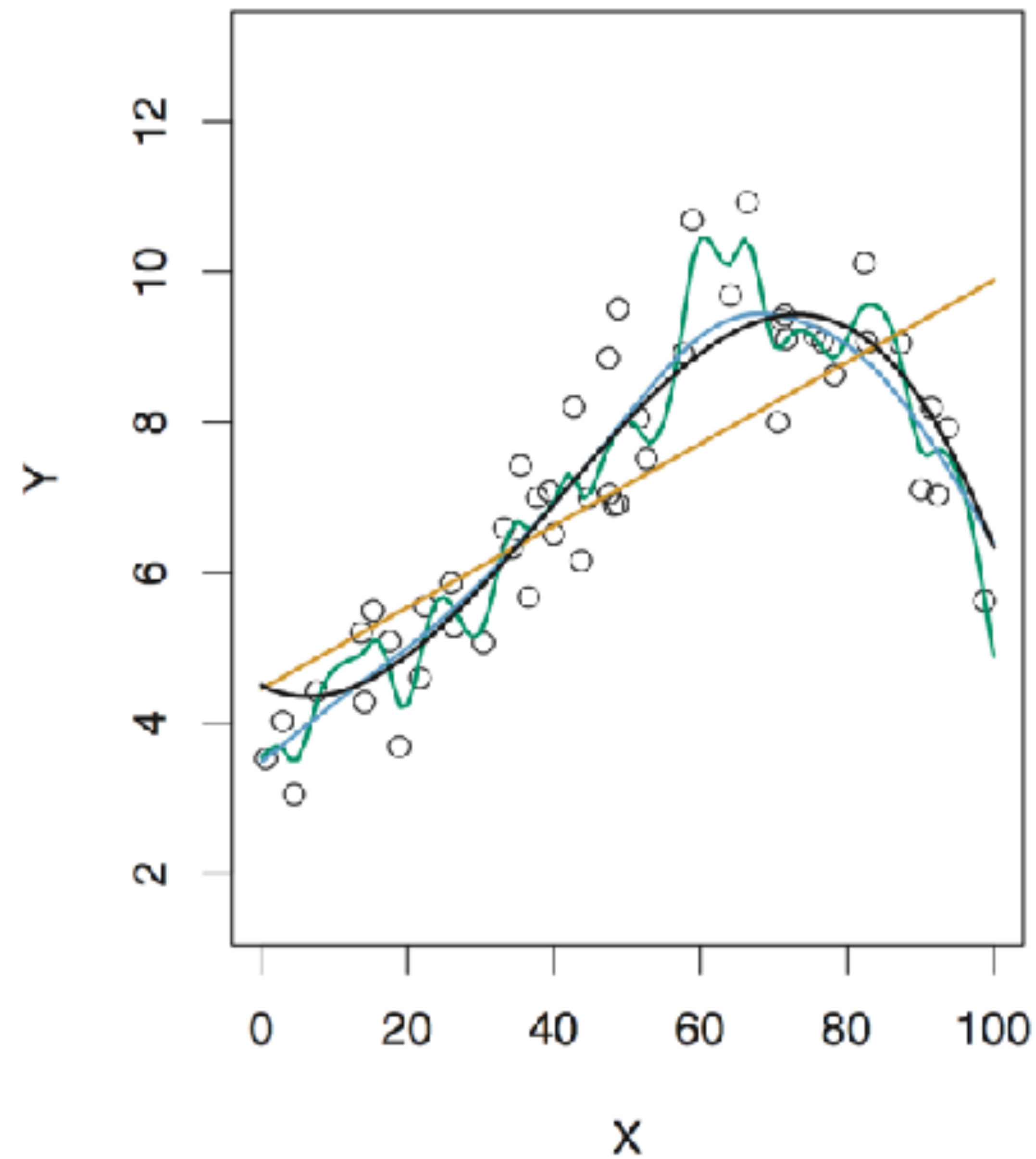
How can we change how well we're doing?

- If making decisions like choosing hyperparameters based on the results of CV runs…
  - Those hyperparameters are chosen to minimize error over all the data they can see (data used within the CV folds)
  - Just as regular-parameters are chosen to minimize error over all the data they can see (data used within a single model)
- Therefore we need a *new*, *completely unseen* set of data on which to evaluate our model's performance
  - We train a model with the optimally-chosen hyperparameter, using all data from our CV folds as training data
  - We use a *new*, *only-just-now-for-the-very-first-time-seen* set of held-out data as our test data
- This means that, before we begin any model creation, if we anticipate having hyperparameters (almost always), and want to ultimately be able to get an honest sense of how we're doing (almost always), we need to set aside a set of data that we don't use until the very end.
  - What we previously referred to as merely Train/Test becomes in the hyperparameter world:
  - Train, Validation, Test

How can we change how well we're doing?

Hyperparameter: degree of polynomial



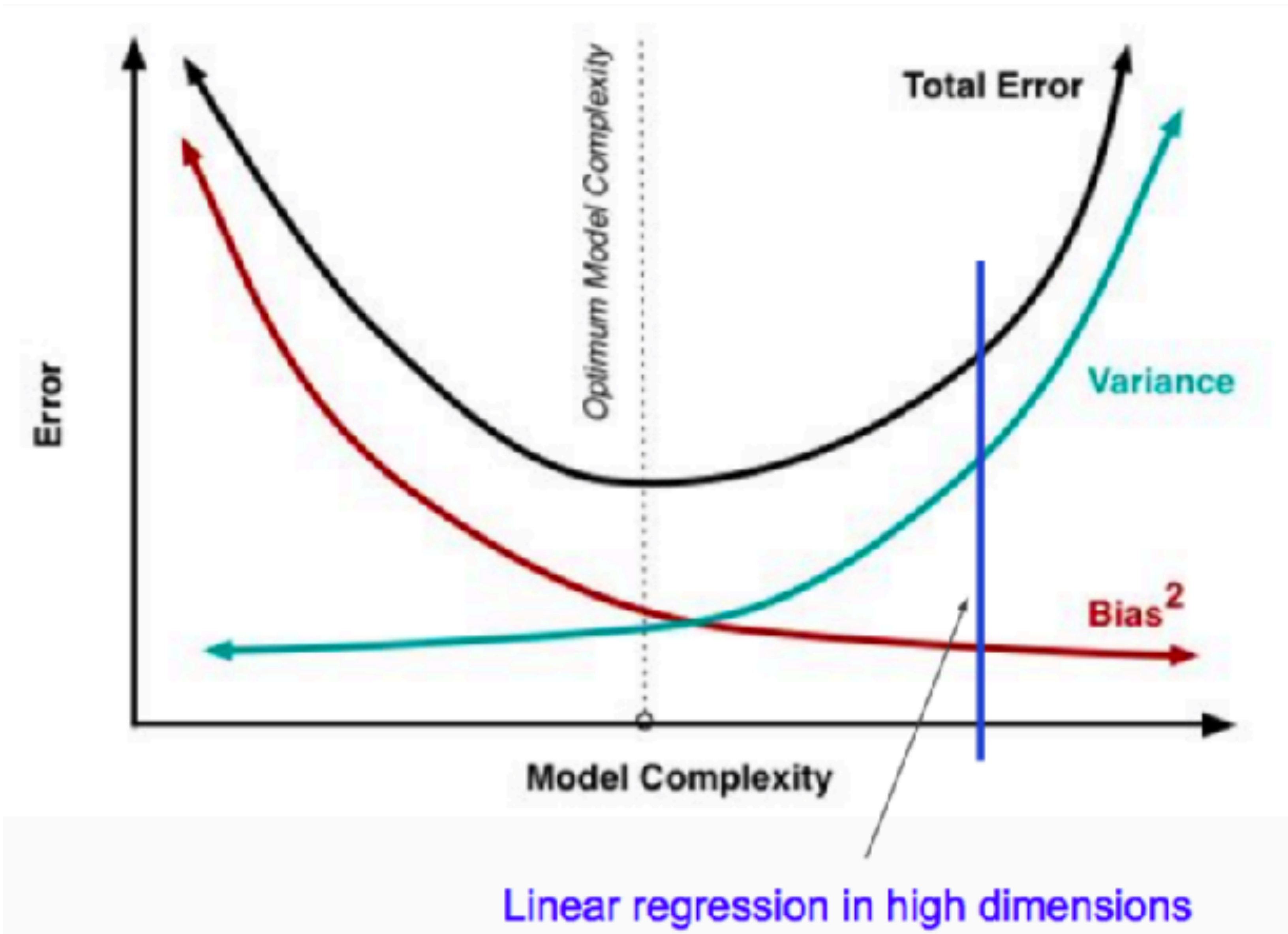


$y = \beta_0 + \beta_1 x_1$

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2$

$y = \beta_0 + \beta_1 x_1 + \beta_1 x_1^2 + \beta_2 x_1^3 + \ldots$

Bias Variance Tradeoff



$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon)$$

expected test MSE

variance: amount f() would change if trained on a different dataset

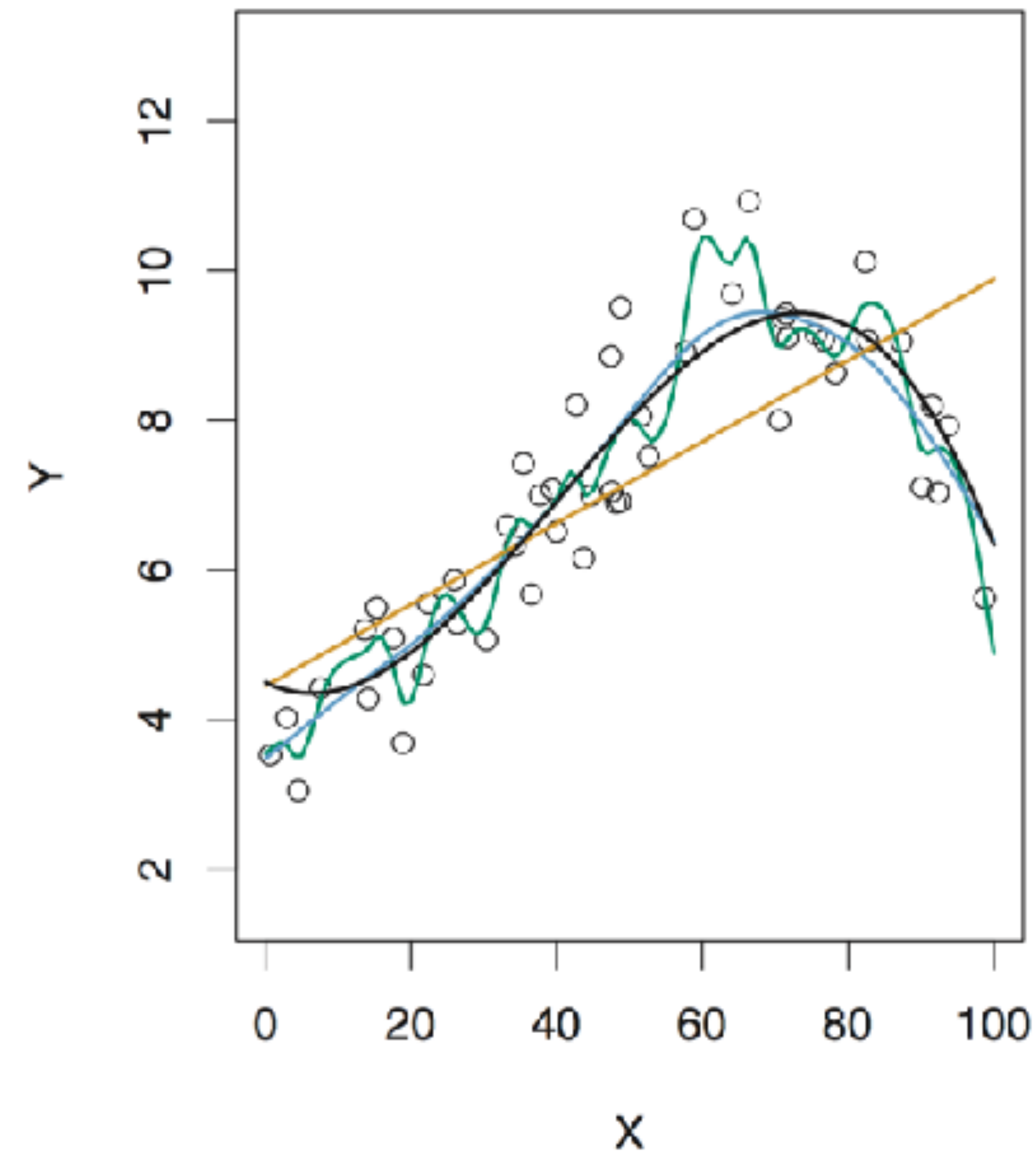
bias: error due to simplified approximation

irreducible error

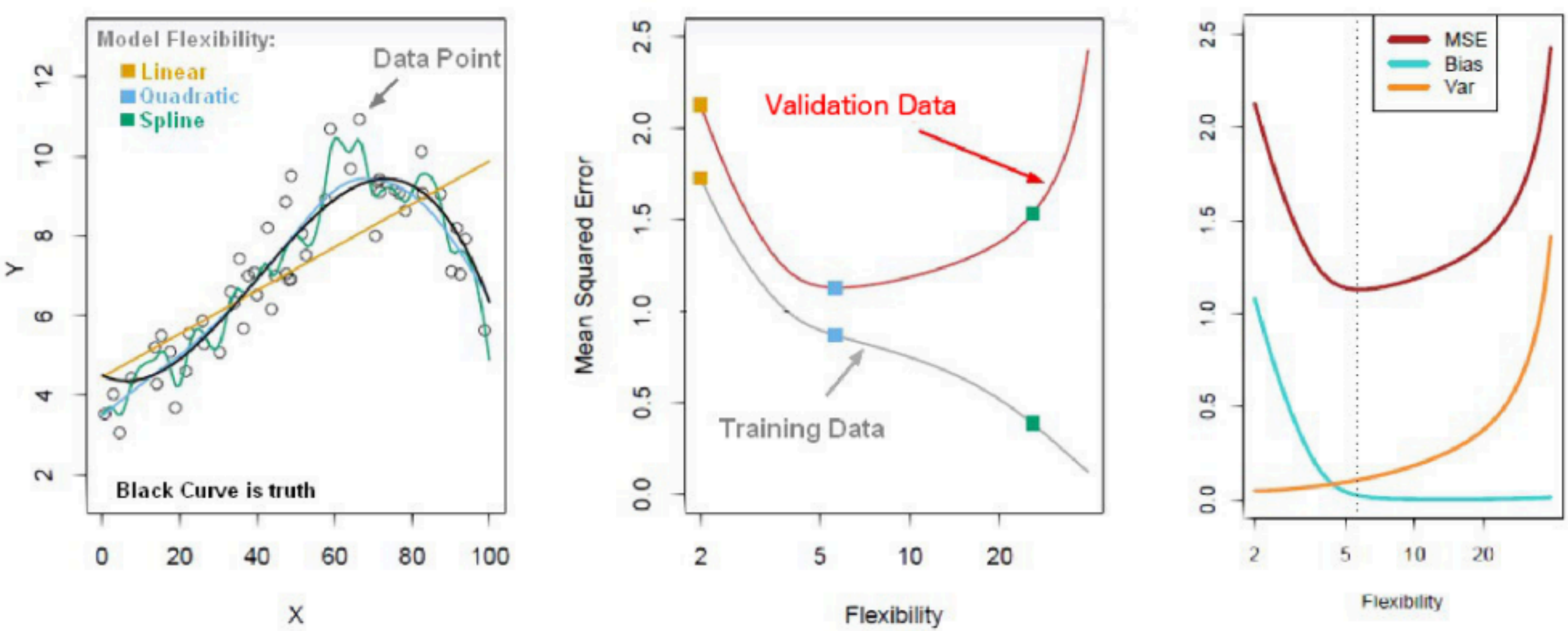
How can we change how well we're doing?

Hyperparameter: degree of polynomial

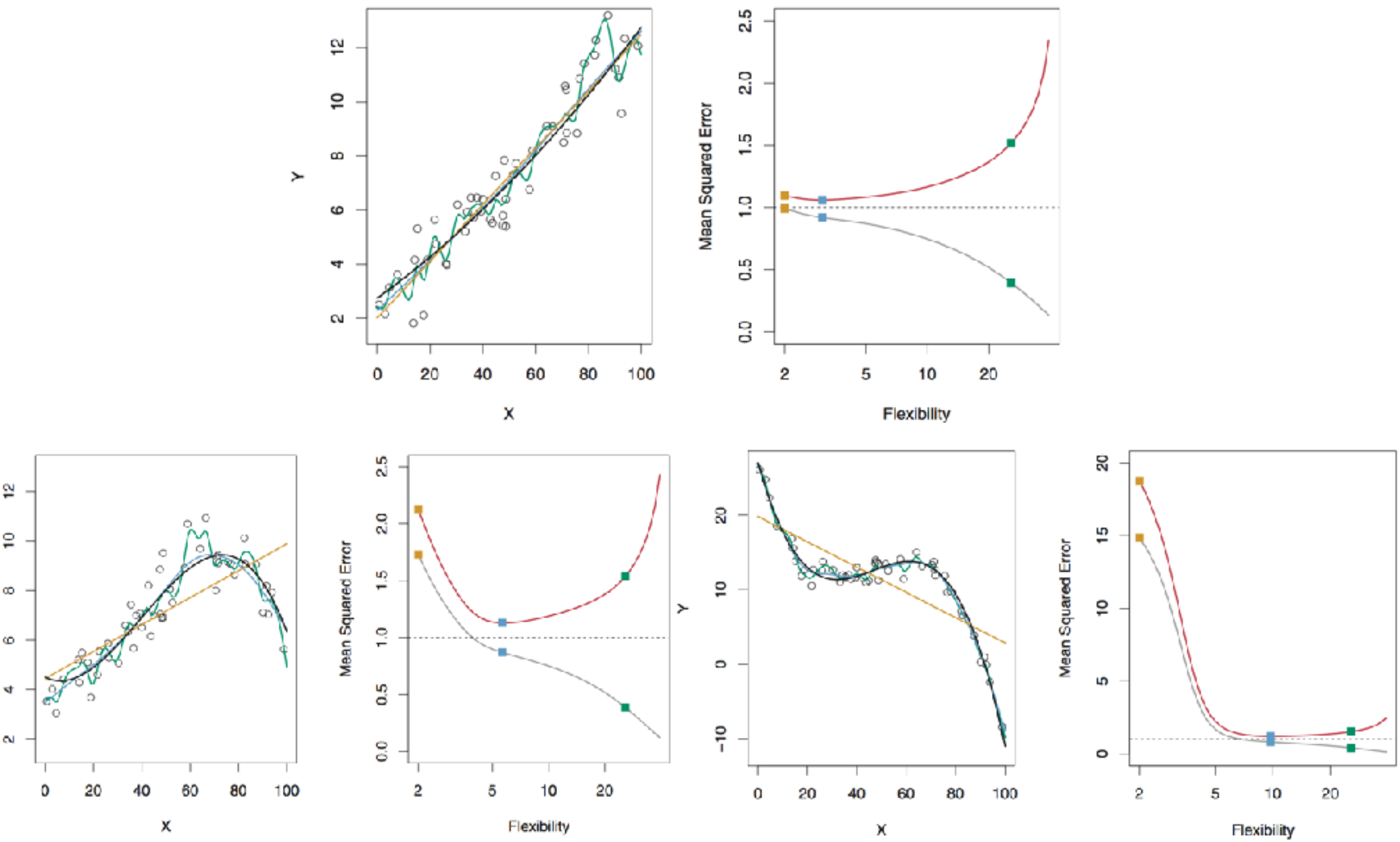
- Underfitting
    - Model does not fully capture the signal in X
    - Insufficiently flexible model
- Overfitting
    - Model erroneously interprets noise as signal
    - Overly flexible model



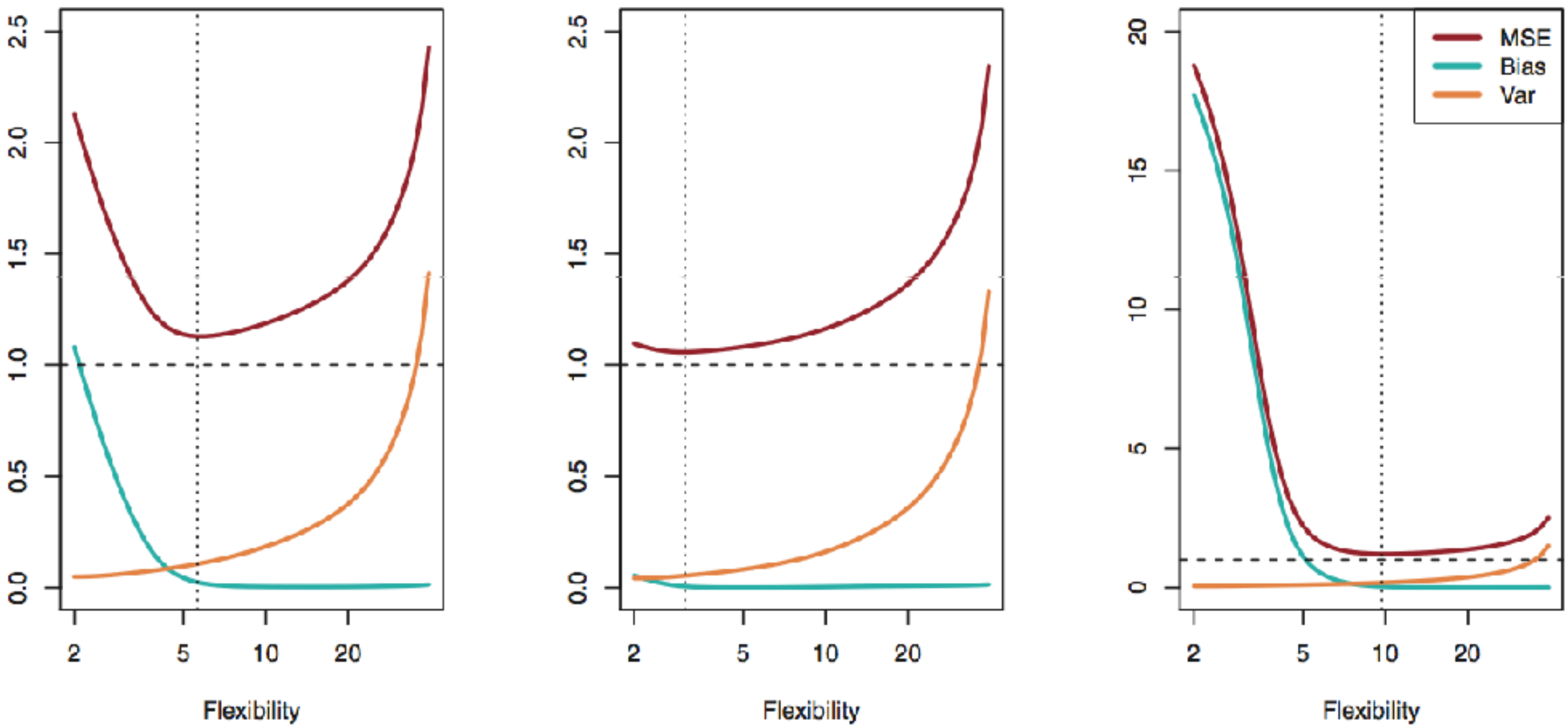
Bias Variance Tradeoff



Bias Variance Tradeoff

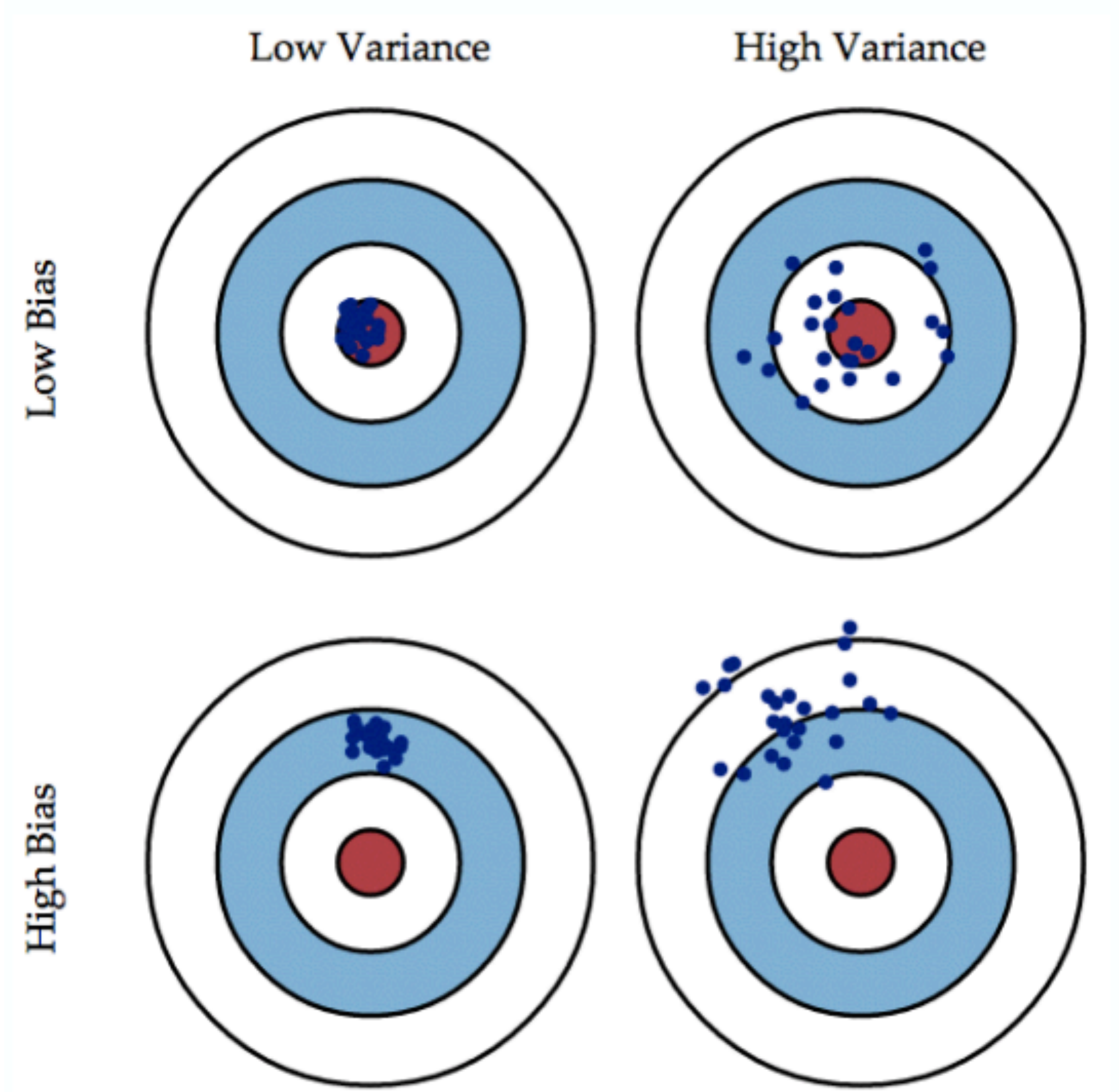


Bias Variance Tradeoff



Caveat: this is pretty contentious since it looks a lot like precision and accuracy, which it's not intended to be

Bias Variance Tradeoff



Goals

- Be able to…
  - explain ridge regression and lasso regression
  - tune the bias/variance of a regression model
  - choose the best regularization hyperparameter for regression

Issues with Ordinary Linear Regression

- High dimensions -> high variance
  - High variance -> overfitting ->   :(…
  - And yet we may want to include dimensions/ features/interactions, if they’re helpful

Ordinary Linear Regression

Find betas to minimize RSS:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2$$

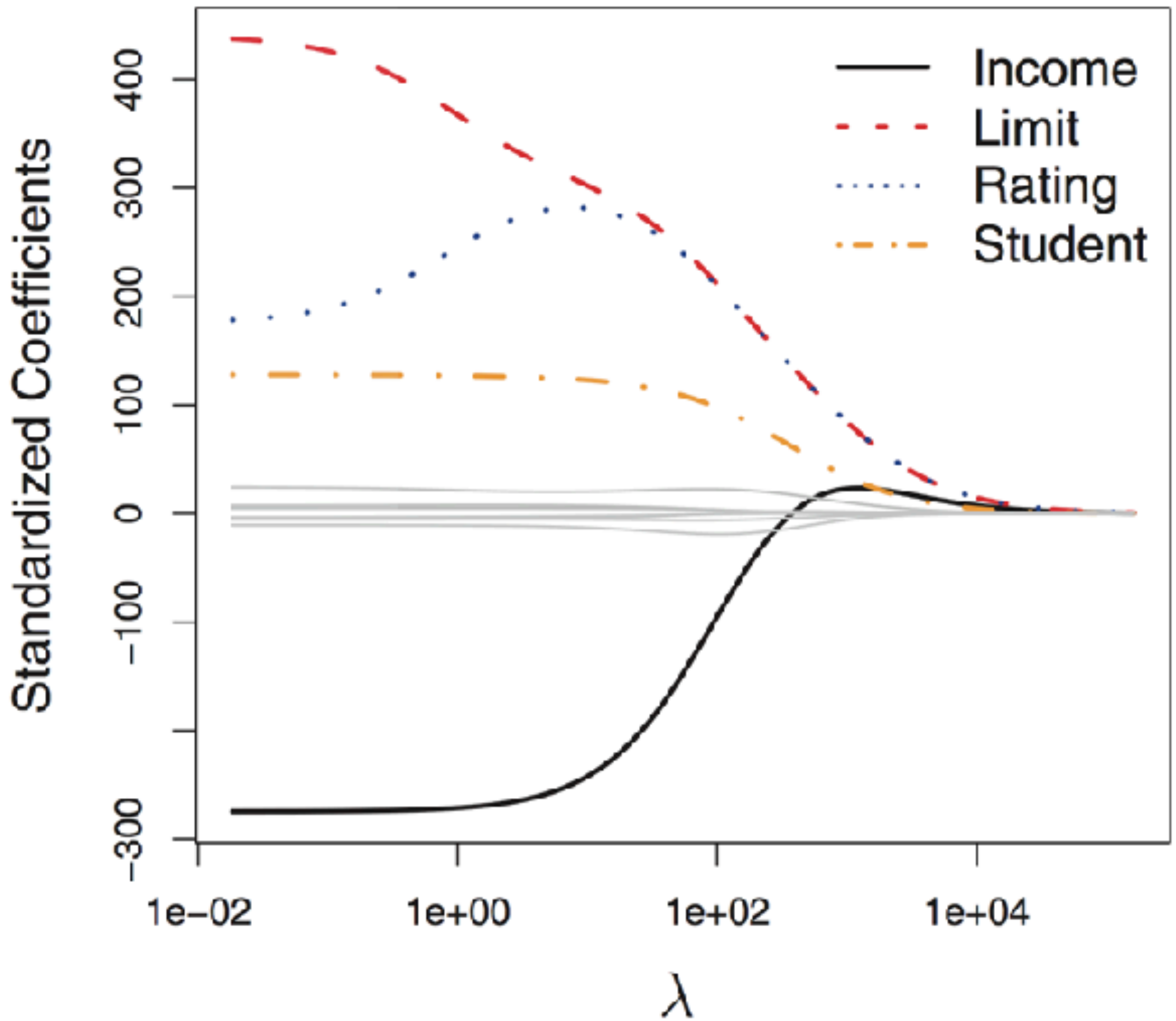
Ridge Regression

Find betas to minimize RSS:

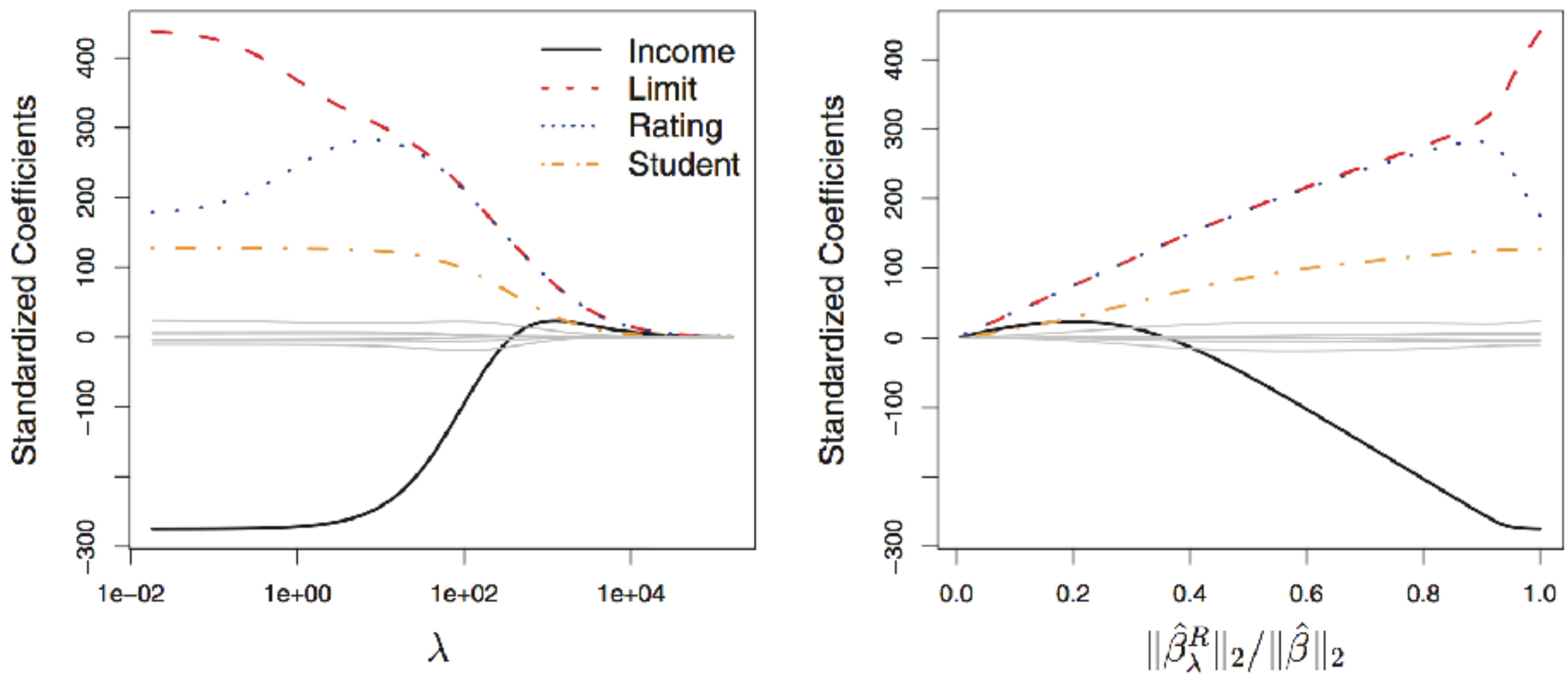
$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda \sum_{j=1}^{p}\beta_j^2$$

your friend, the hyperparameter

Ridge Regression



Ridge Regression



l2 ("ell two") norm: $\|\beta\|_2 = \sqrt{\sum_{j=1}^{p} {\beta_j}^2}$

x axis on the right: amount that the ridge regression coefficient estimates have been shrunken towards zero; a small value indicates that they have been shrunken very close to zero

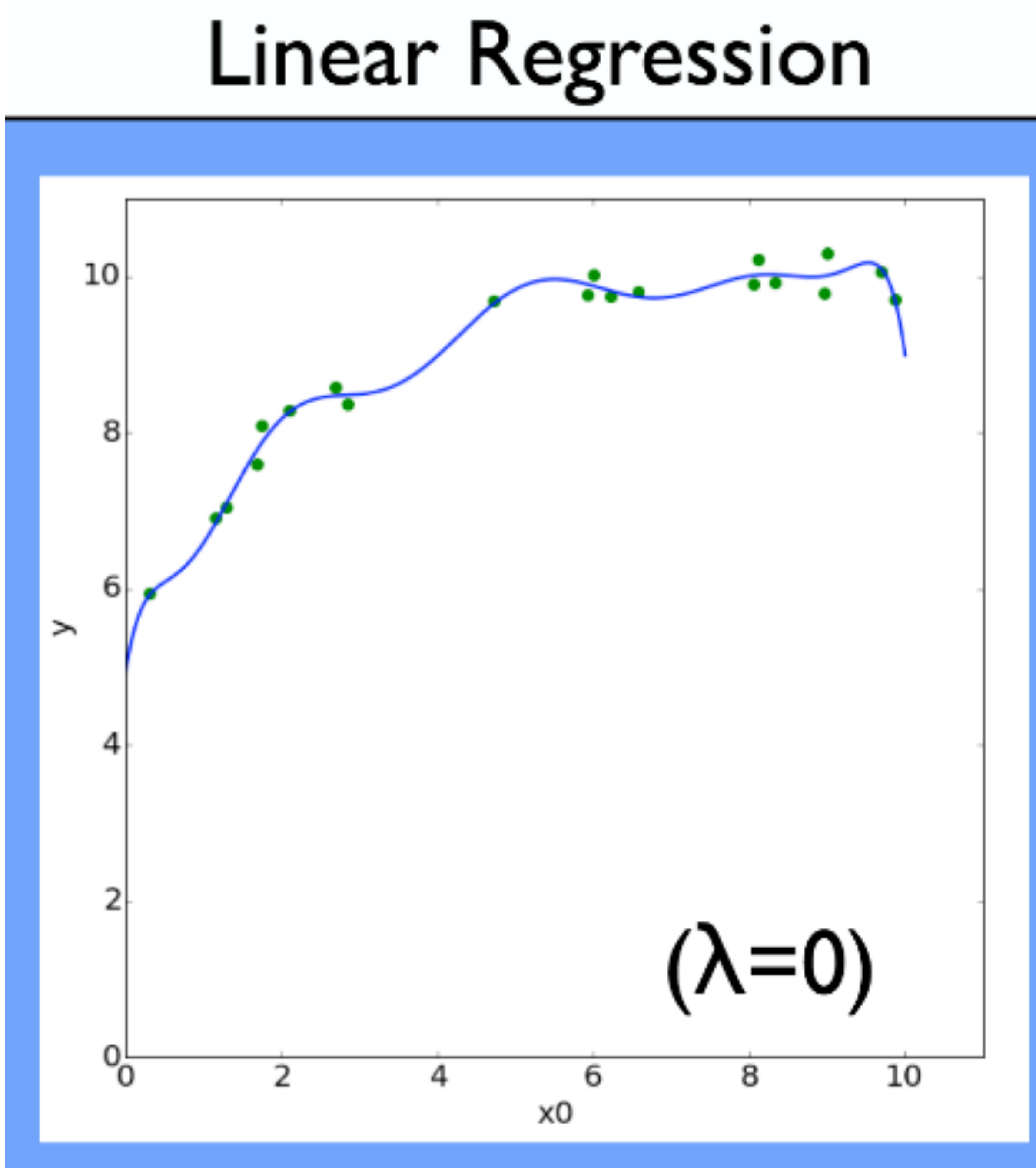
Ridge Regression

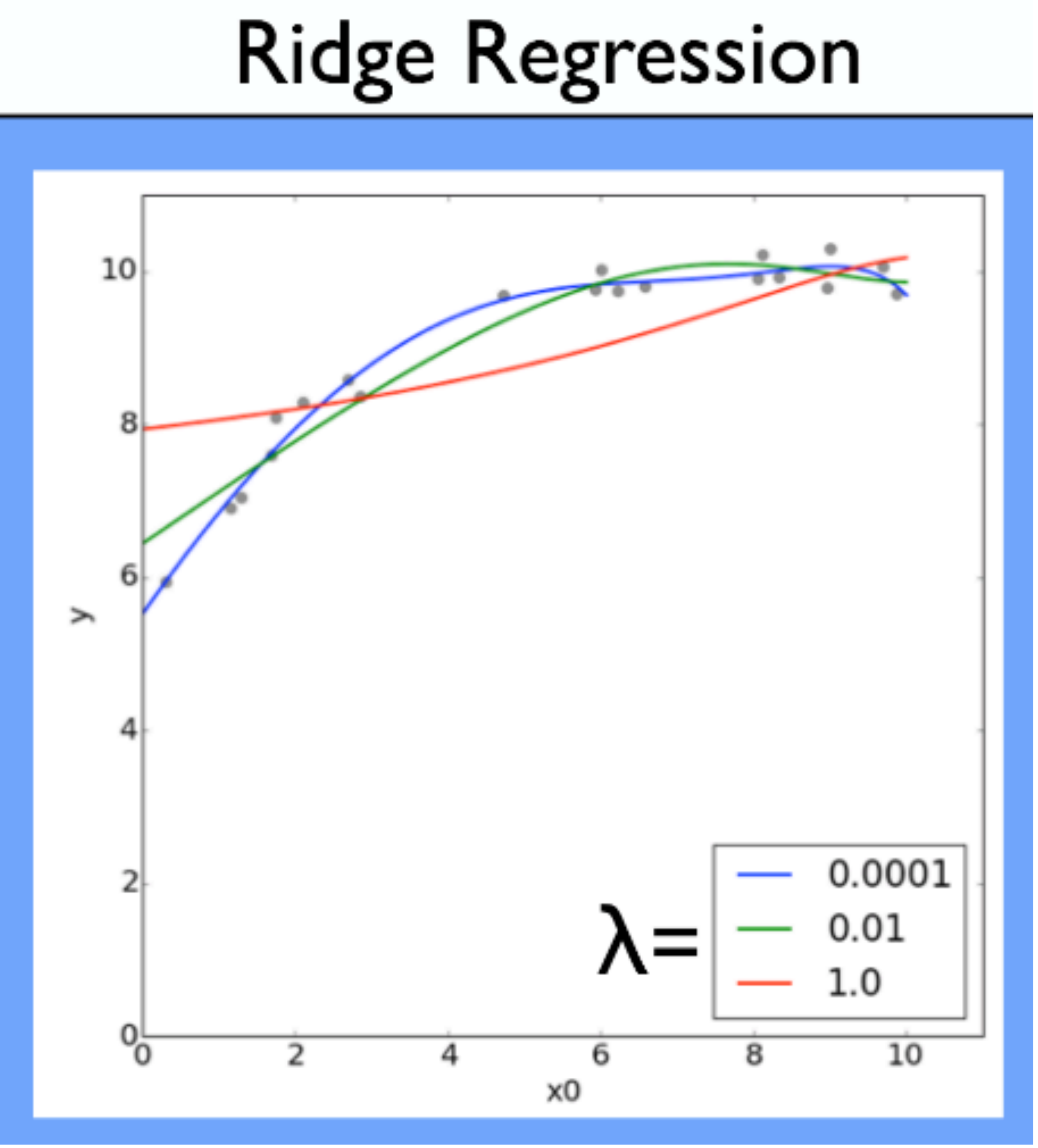
Warning: when using ridge regression, scale matters!
Why? (Consider units to measure salary)

Standardize your predictors:

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2}}$$

Ridge Regression





Lasso Regression

Find betas to minimize RSS:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

your friend, the hyperparameter

remember, ridge:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

Ridge and Lasso

Ridge:

minimize RSS:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

l2 (“ell two”) norm:

$$\|\beta\|_2 = \sqrt{\sum_{j=1}^{p}\beta_j{}^2}$$

Lasso:

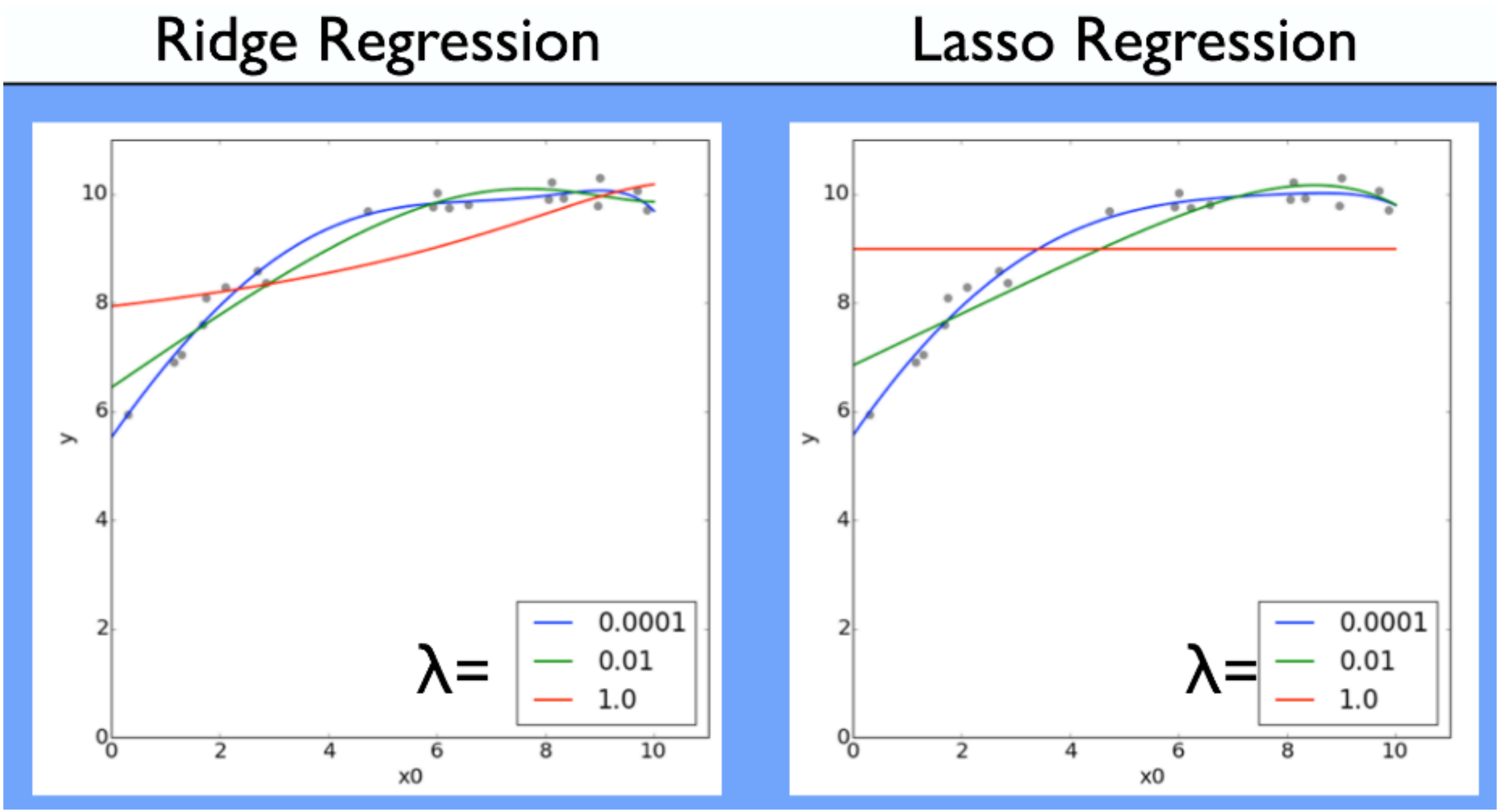
minimize RSS:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

l1 (“ell one”) norm:

$$\|\beta\|_1 = \sum|\beta_j|$$

Ridge and Lasso



Ridge vs. Lasso