Ridge and Lasso - Hyperparameters

CS1090A Introduction to Data Science

Pavlos Protopapas, Kevin Rader, and Chris Gumb

Photo: Laine Roper
Kings Canyon, California

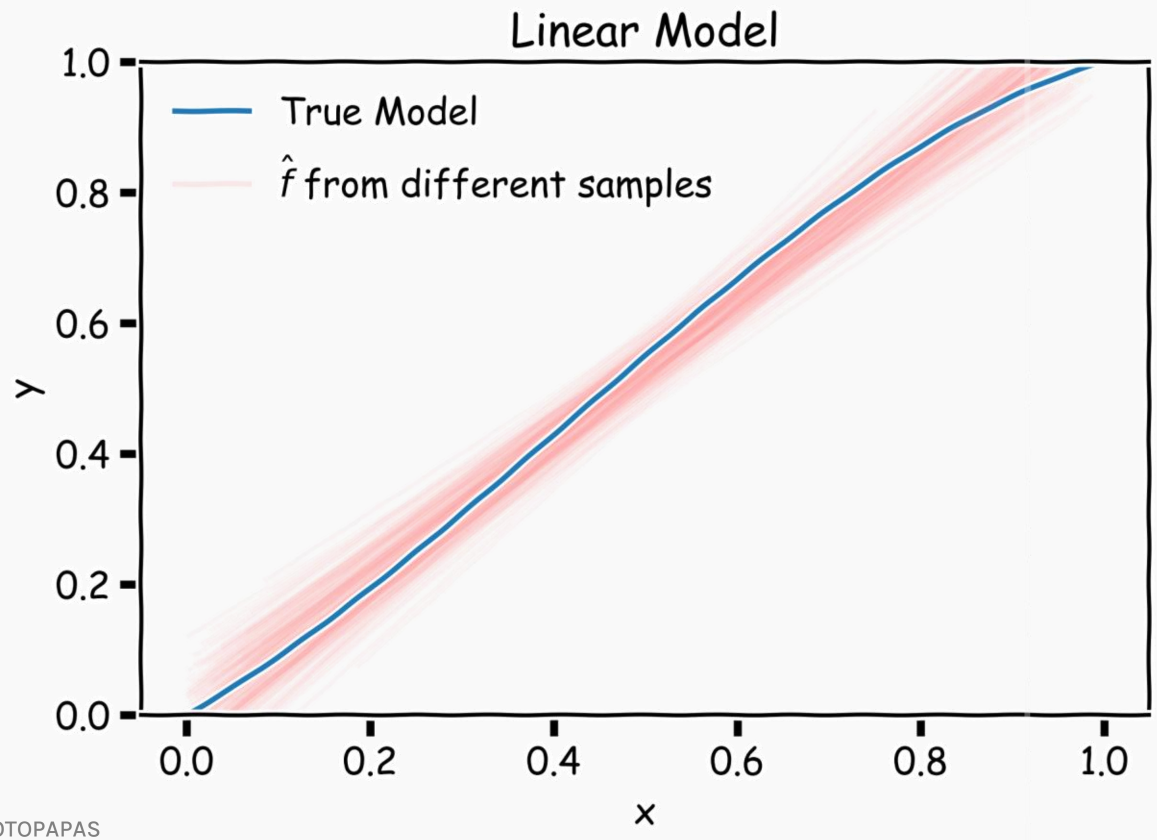
Outline

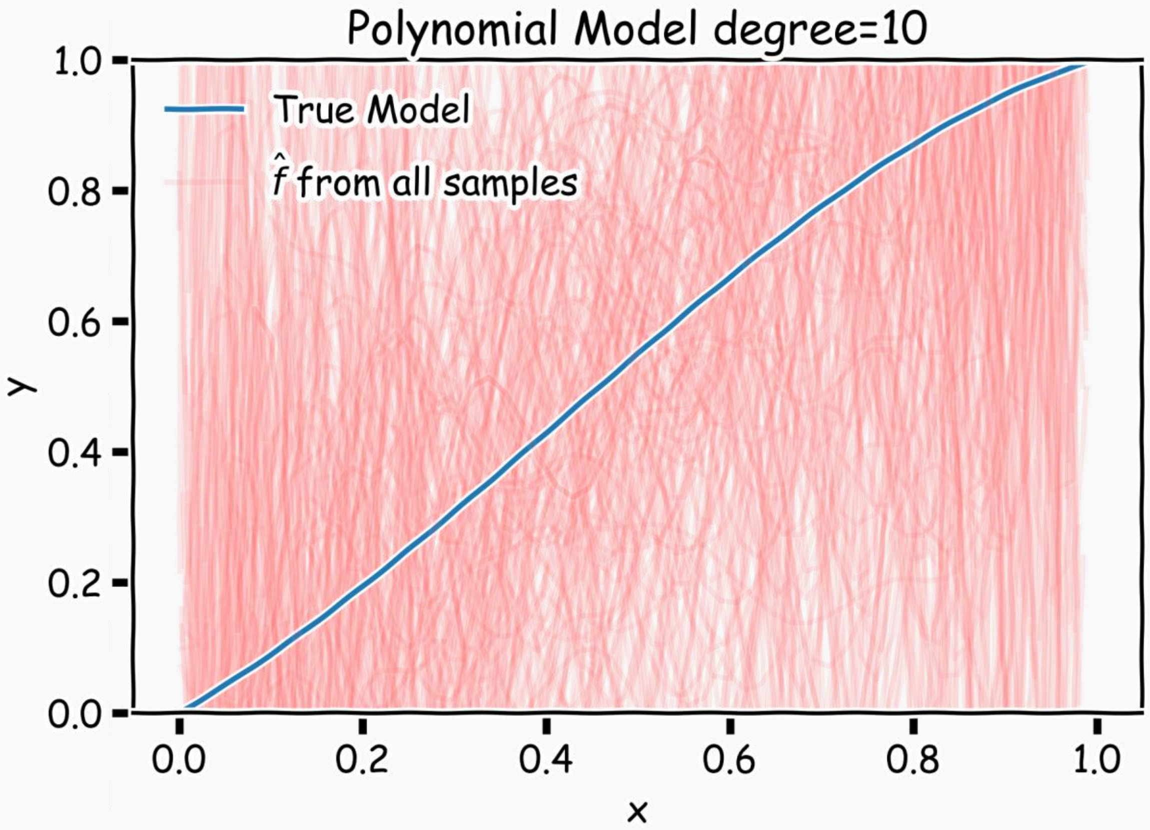
- Recap – Model Selection
- Generalization Error, Bias Variance Tradeoff
- **Regularization Techniques: Lasso, Ridge**

Bias vs Variance

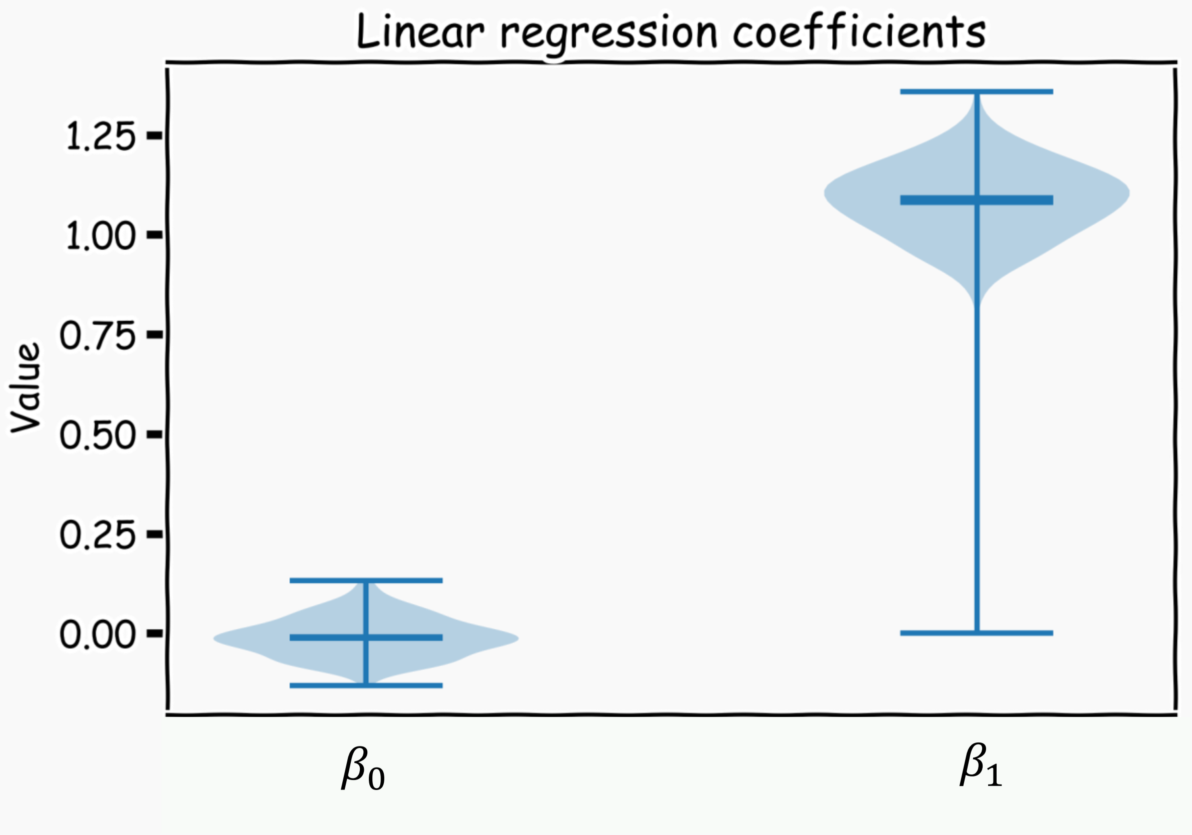
Left: 2000, best fit straight lines, each fitted on a different 20-point training set.

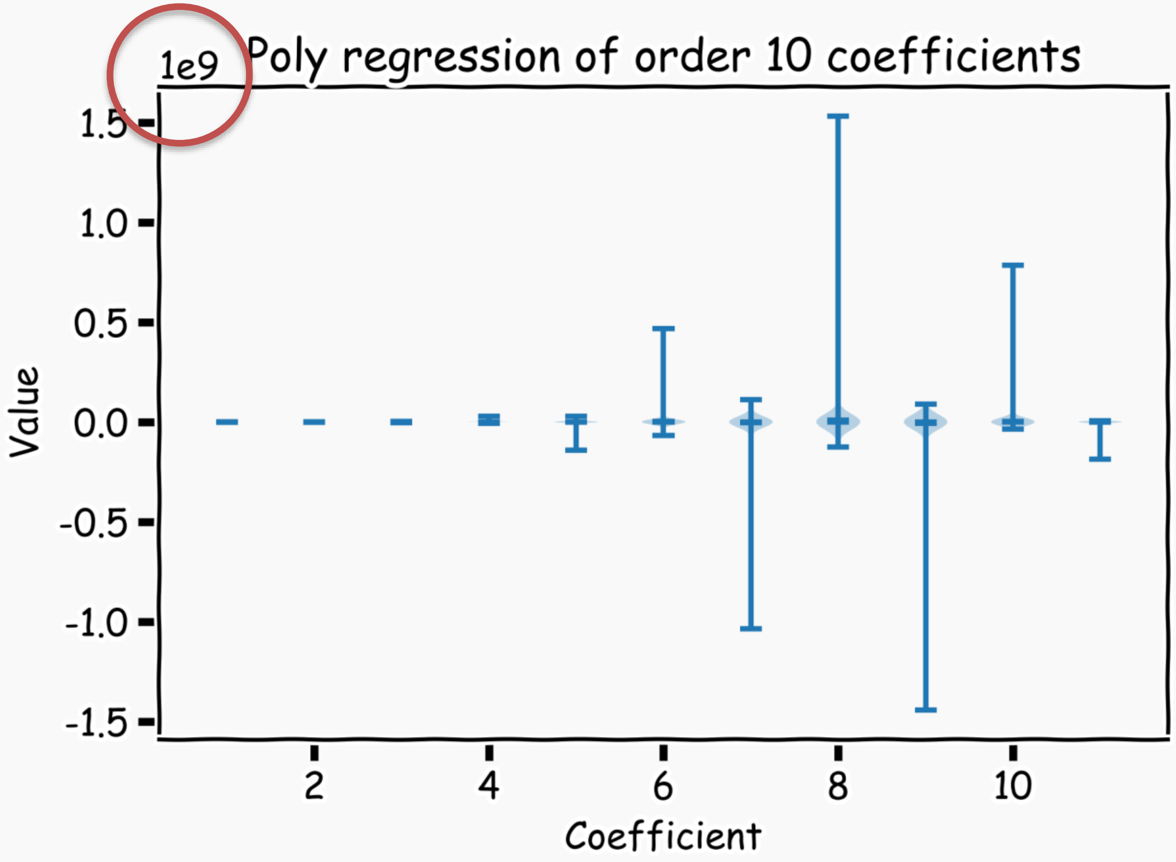
Right: Best-fit models using degree-10 polynomial





Bias vs Variance





Model Selection

?

Model selection is the application of a principled method to determine the complexity of the model, e.g., choosing a subset of predictors, choosing the degree of the polynomial model etc.

A strong m oid overfitting,

How do we discourage extreme values in the model parameters?

- there are
  - the
  - the polynomial degree is too high
  - too many cross terms are considered
- the coefficients values are too **extreme**

Game Time

How would you discourage extreme values in the model parameters

Options:

A. Divide all model parameters by a large number

B. Make sure the causal relationship between predictors and response variable is true

C. Discard any model with model parameter value larger than 1

D. Penalize the model with a penalty that is proportional to the value its parameters

Regularization

What we want

Low model error

Minimize:

$$\frac{1}{n}\sum_{i=1}^{n}\left|y_i - \boldsymbol{\beta}^\top \boldsymbol{x}_i\right|^2$$

Discourage extreme values in model parameters

Minimize:

Regularization

What we want

Low model error

Minimize:

$$\frac{1}{n}\sum_{i=1}^{n}\left|y_i - \boldsymbol{\beta}^\top \boldsymbol{x}_i\right|^2$$

Discourage extreme values in model parameters

Minimize:

$$L_{reg} = \begin{cases} \sum_{j=1}^{J} \beta_j^2 \\ \sum_{j=1}^{J} |\beta_j| \end{cases}$$

PROTOPAPAS

Regularization

What we want

Low model error

Minimize:

$$\frac{1}{n}\sum_{i=1}^{n}\left|y_i - \boldsymbol{\beta}^\top \boldsymbol{x}_i\right|^2$$

Discourage extreme values in model parameters

Minimize:

$$L_{reg} = \begin{cases} \sum_{j=1}^{J} \beta_j^2 \\ \sum_{j=1}^{J} |\beta_j| \end{cases}$$

How do we combine these two objectives?

Regularization

What we want

Regularization

What we want

Low model error

Minimize:

Discourage extreme values in model parameters

Minimize:

$$\mathcal{L}_{REG} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \boldsymbol{\beta}^\top \boldsymbol{x}_i\right|^2 + L_{reg}$$

Regularization

What we want

Low model error

Minimiz

Discourage extreme values in model parameters

mize:

λ is the **regularization parameter**. It controls the relative importance between model error and the regularization term

$$\mathcal{L}_{REG} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \boldsymbol{\beta}^\top \boldsymbol{x}_i\right|^2 + \lambda\, L_{reg}$$

Regularization

What we want

Low model error

Discourage extreme values in model parameters

$\lambda = 0$: equivalent to simple linear regression

$\lambda = \infty$: yields a model with $\beta's$ =0

mize:

$$\mathcal{L}_{REG} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \boldsymbol{\beta}^\top \boldsymbol{x}_i\right|^2 + \lambda L_{reg}$$

Regularization

What we want

Low model error

Minimize:

Discourage extreme values in model parameters

Minimize:

How do we determine λ?

$$\mathcal{L}_{REG} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \boldsymbol{\beta}^\top \boldsymbol{x}_i\right|^2 + \lambda L_{reg}$$

Regularization

What we want

Low model error

Minimize:

Discourage extreme values in model parameters

Minimize:

Cross Validation!

$$\mathcal{L}_{REG} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \boldsymbol{\beta}^\top \boldsymbol{x}_i\right|^2 + \lambda\, L_{reg}$$

Regularization: LASSO Regression

What we want

Low model error

Minimize:

Discourage extreme values in model parameters

Minimize:

Note that $\sum_{j=1}^{J} |\beta_j|$ is the l_1 norm of the vector $\boldsymbol{\beta}$

$$\mathcal{L}_{LASSO} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \boldsymbol{\beta}^\top \boldsymbol{x}_i\right|^2 + \lambda \sum_{j=1}^{J} |\beta_j|$$

Regularization: **LASSO** Regression

What we want

Low model error

Discourage extreme values in
...l parameters

...nize:

No need to regularize the bias, β_0
Why?

$$\mathcal{L}_{LASSO} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \boldsymbol{\beta}^\top \boldsymbol{x}_i\right|^2 + \lambda \sum_{j=1}^{J}|\beta_j|$$

Regularization: **LASSO** Regression

Lasso regression: minimize $\mathcal{L}_{LASSO}$ with respect to $\beta's$

$$\mathcal{L}_{LASSO} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \boldsymbol{\beta}^\top \boldsymbol{x}_i\right|^2 + \lambda\sum_{j=1}^{J}|\beta_j|$$

Ridge regression: minimize $\mathcal{L}_{RIDGE}$ with re

Note that $\sum_{j=1}^{J} \beta_j^2$ is the $\boldsymbol{L_2}$ norm square of the vector $\boldsymbol{\beta}$

$$\mathcal{L}_{RIDGE} = \frac{1}{n} \sum_{i=1}^{n} \left| y_i - \boldsymbol{\beta}^\top \boldsymbol{x}_i \right|^2 + \lambda \sum_{j=1}^{J} \beta_j^2$$

Regularization: Ridge Regression

Ridge regression: minimize $\mathcal{L}_{RIDGE}$ with respect to $\beta's$

$$\mathcal{L}_{RIDGE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \boldsymbol{\beta}^\top \boldsymbol{x}_i\right|^2 + \lambda \sum_{j=1}^{J}\beta_j^2$$

No need to regularize the bias, β_0, since it is not connected to the predictors.

Ridge regularization with only **validation** : step by step

For ridge regression there exist an analytical solution for the coefficients:

$$\hat{\beta}_{Ridge}(\lambda) = (\mathrm{X}^{\mathrm{T}}\mathrm{X} + \lambda I)^{-1} X^T Y$$

So, what are the steps to determine λ using validation?

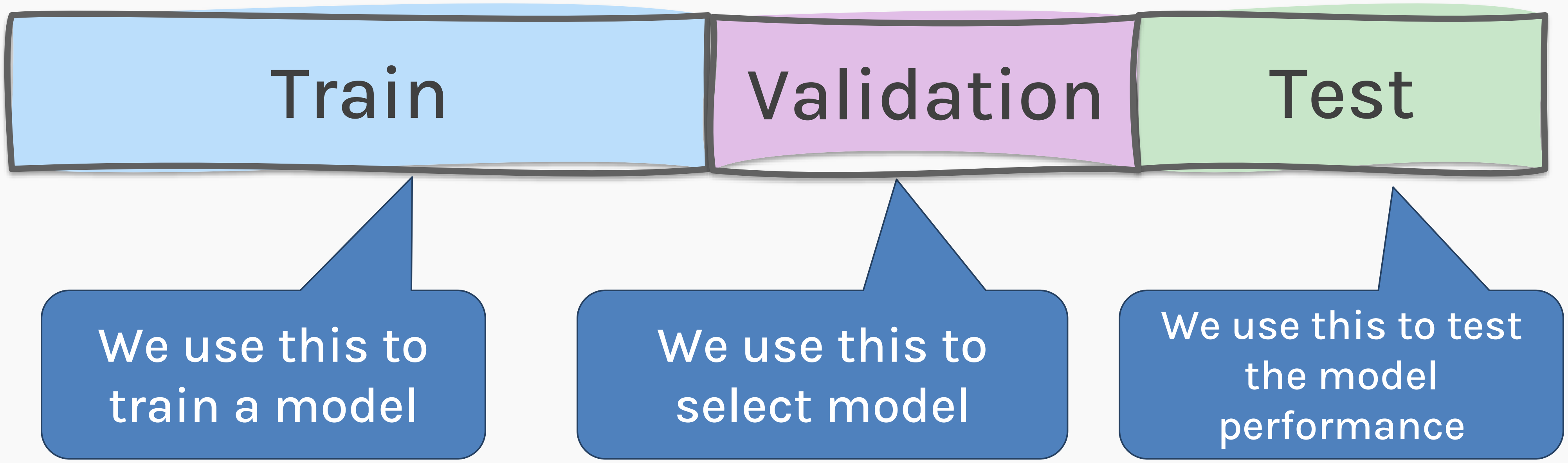
Ridge regularization with only **validation** : step by step

For ridge regression there exist an analytical solution for the coefficients:

$$\hat{\beta}_{Ridge}(\lambda) = \left(\mathrm{X}^{\mathrm{T}}\mathrm{X} + \lambda I\right)^{-1} X^T Y$$

Step 1 : Split Data

Split data into train, validation and test data.



Ridge regularization with only **validation** : step by step

Step 2 : Select a range of possible λ values

λ_{min}	λ_{min-1}	….	λ_{max-1}	λ_{max}

The values of λ are not fixed—you can choose the range yourself based on the problem.

Ridge regularization with only **validation** : step by step

Step 2 : Select a range of possible λ values

As an example, we will take a range of values from 0.00001 to 10

$\lambda\ values =$

0.00001	0.0001	0.001	0.01	0.1	....	1	2	3	4	10

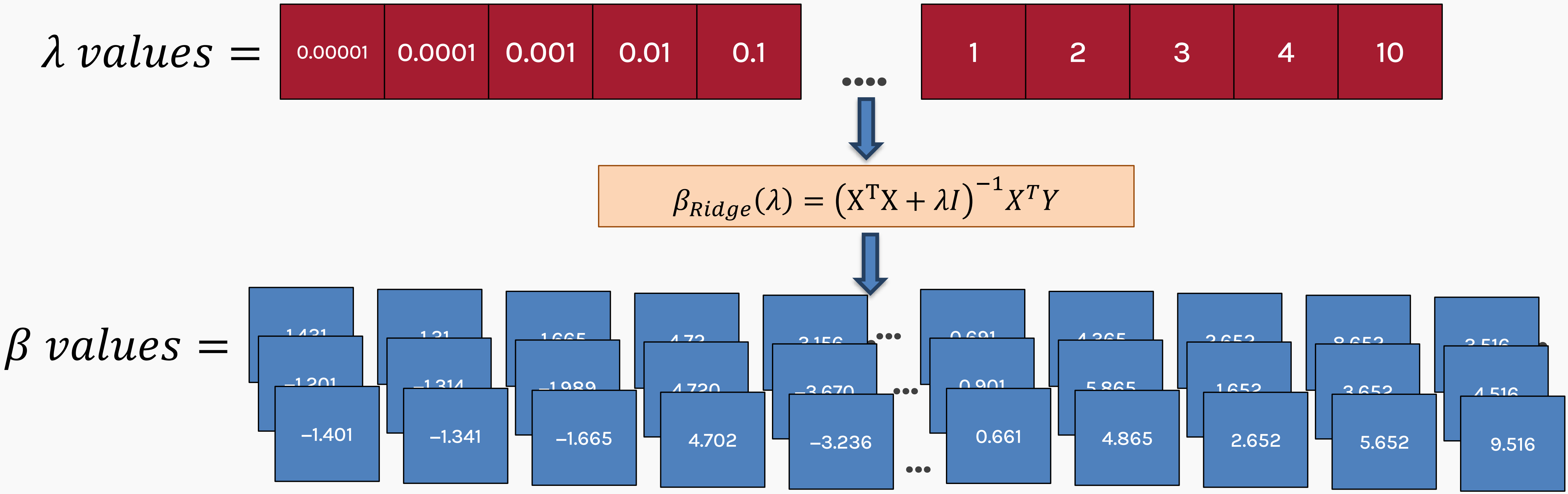
λ_{min} λ_{max}

Ridge regularization with only **validation** : step by step

Step 3 : Determine β

Train

For each λ value, we determine $\beta_{Ridge}(\lambda)$ using train data



PROTOPAPAS

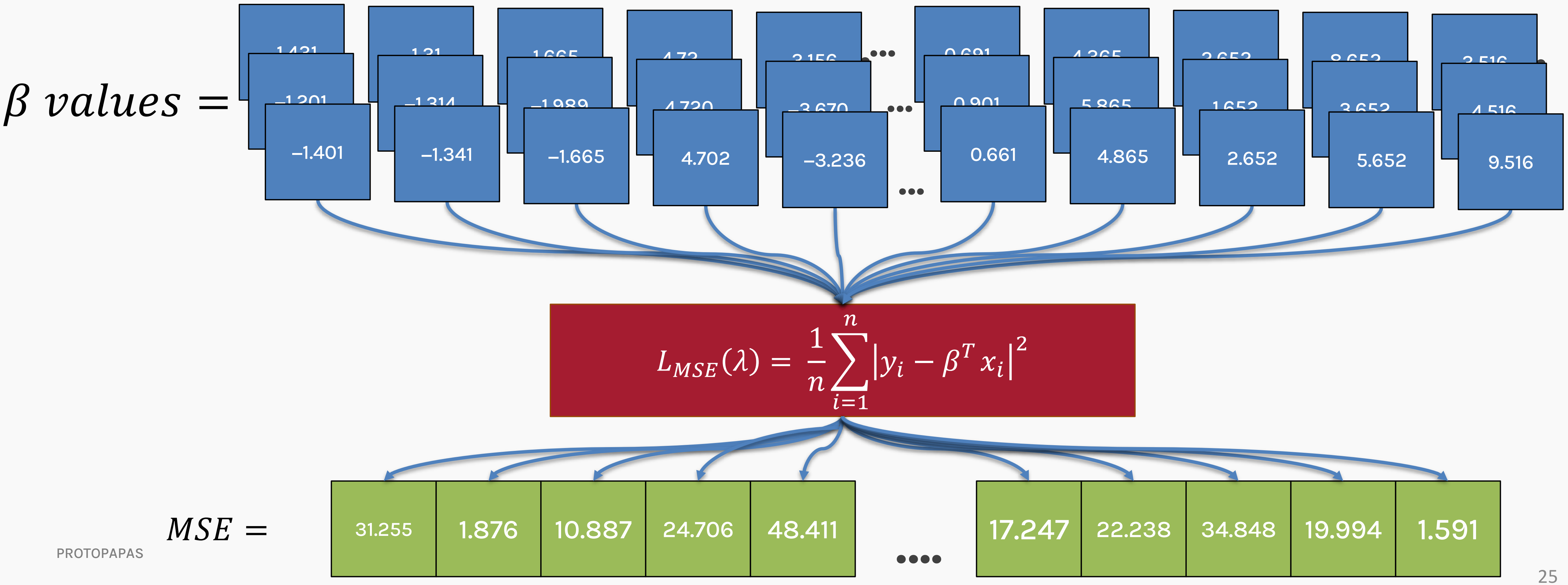
Keep in mind that each λ value gives us a different vector of β coefficients.

Ridge regularization with only **validation** : step by step

Step 4 : Record the $L_{MSE}(\lambda)$

Validation

For each vector of β coefficients, we record $L_{MSE}(\lambda)$ using the validation data



Ridge regularization with only **validation** : step by step

Step 5 : Select the λ_{Ridge}

Validation

Select the λ that minimizes the MSE loss on the validation data.

$\lambda\ values =$

0.00001	0.0001	0.001	0.01	0.1	....	1	2	3	4	10

$MSE\ =$

31.255	1.876	10.887	24.706	48.411	....	17.247	22.238	34.848	19.994	1.591

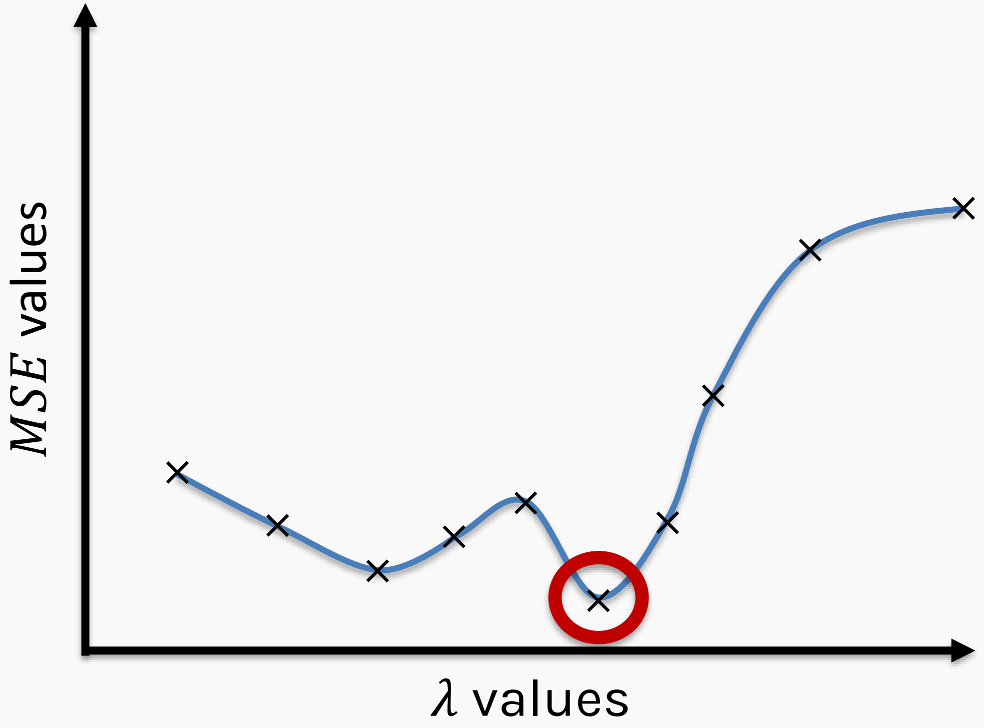
Let's visualize and see it!

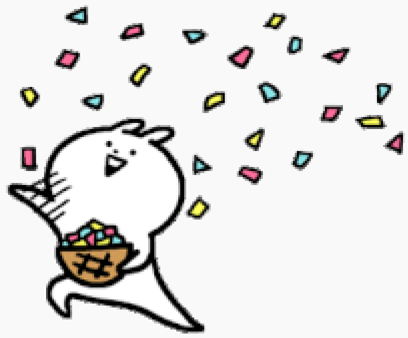
Ridge regularization with only **validation** : step by step

Step 5 : Select the λ_{Ridge}

Validation

Select the λ that minimizes the MSE loss on the validation data.

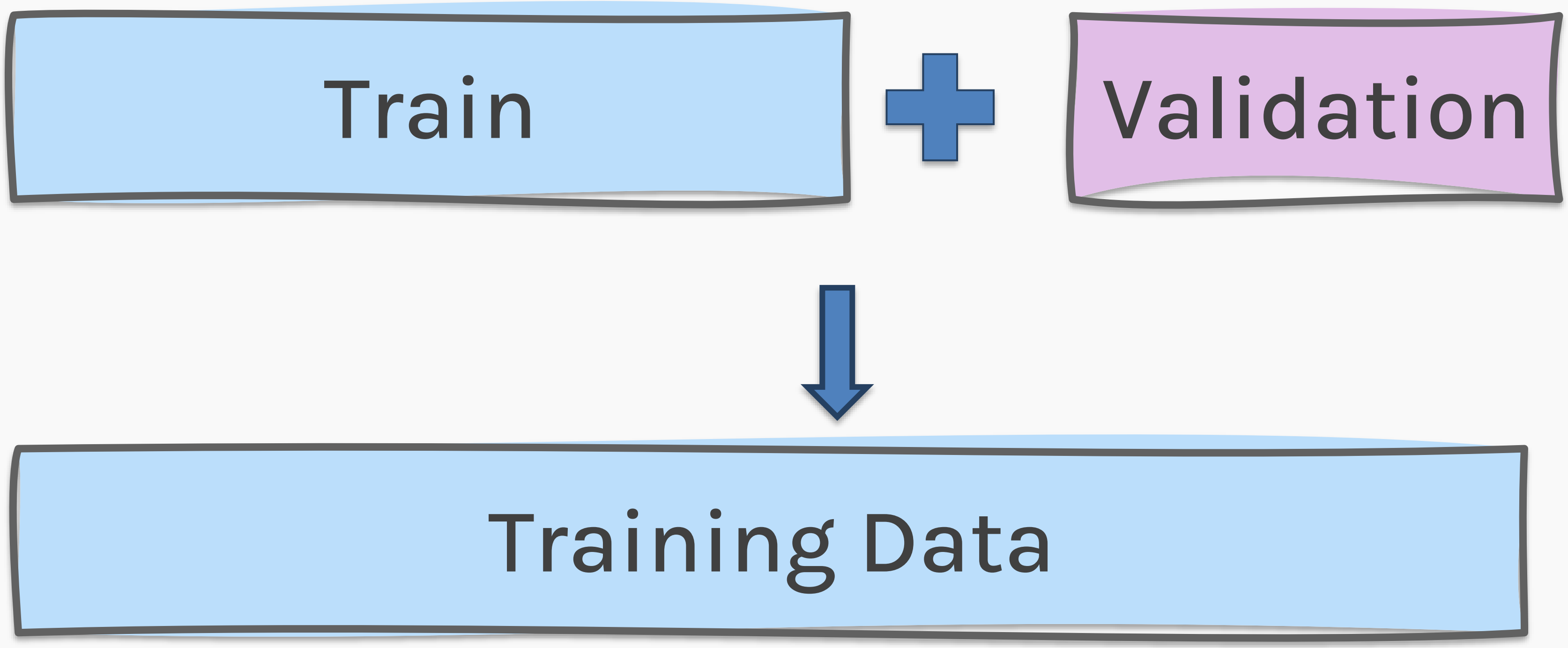




Ridge regularization with only **validation** : step by step

Step 6 : Refit the model

Refit the model using both train and validation data using λ_{Ridge}.

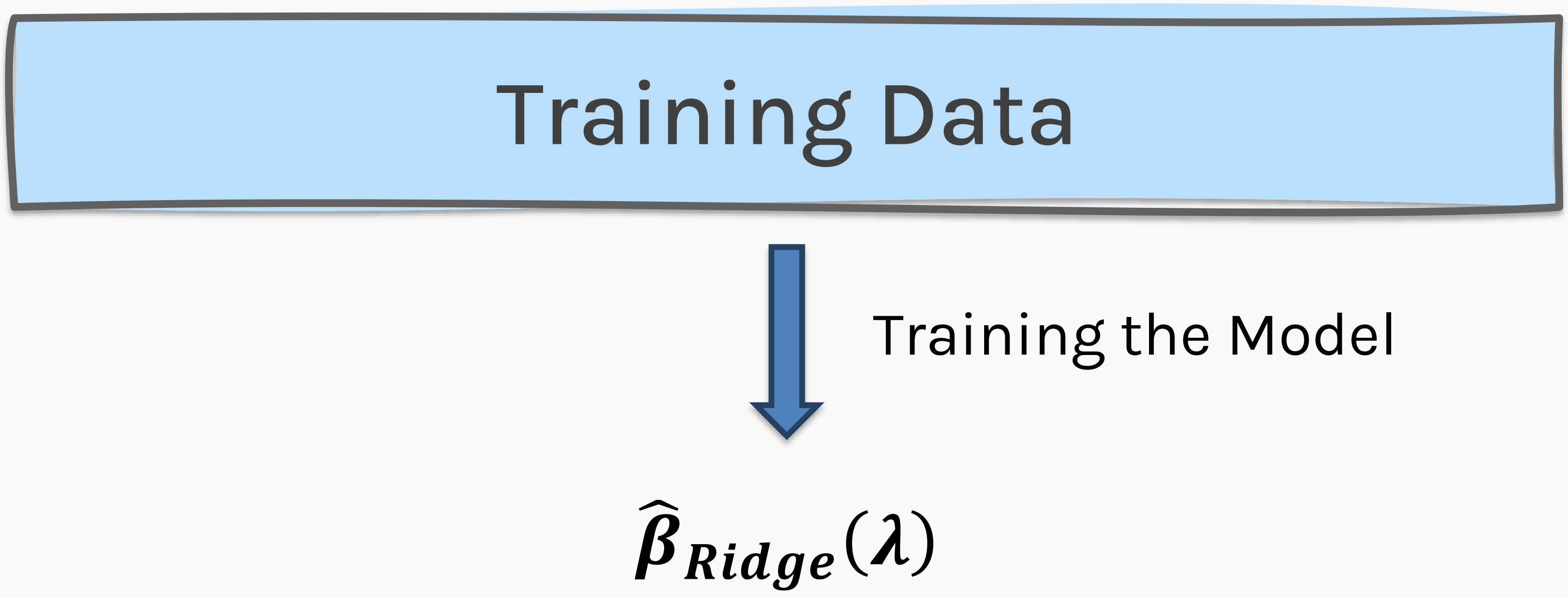


Ridge regularization with only **validation** : step by step

Step 6 : Refit the model

Refit the model using both train and validation data using λ_{Ridge}.

This gives us $\widehat{\beta}_{Ridge}(\lambda)$



Ridge regularization with only **validation** : step by step

Step 7 : Record MSE/R2

Report MSE or R2 on test data given the $\widehat{\boldsymbol{\beta}}_{Ridge}(\lambda)$

Lasso regularization with only **validation** : step by step

For Lasso regression, there is **no** analytical solution for the coefficients, so we use a **solver**.

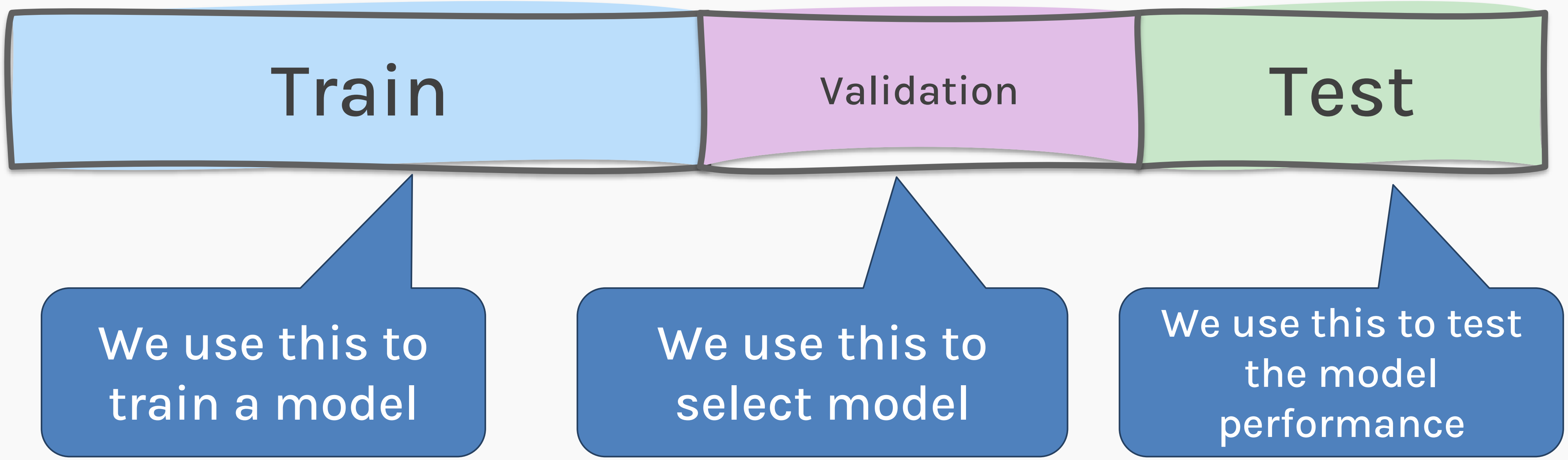
Now, the steps to determine λ using validation will still be the same!

Lasso regularization with only **validation** : step by step

For Lasso regression, there is **no** analytical solution for the coefficients, so we use a **solver**.

Step 1 : Split Data

Split data into train, validation and test data.



Lasso regularization with only **validation** : step by step

Step 2 : Select a range of possible λ values

λ_{min}	λ_{min-1}	....	λ_{max-1}	λ_{max}

The values of λ are not fixed—you can choose the range yourself based on the problem.

Lasso regularization with only **validation** : step by step

Step 2 : Select a range of possible λ values

As an example, we will take a range of values from 0.00001 to 10

$$\lambda\ values = \begin{array}{|c|c|c|c|c|}\hline \underset{\uparrow}{0.00001} & 0.0001 & 0.001 & 0.01 & 0.1 \\ \hline \end{array} \ \cdots\cdots \ \begin{array}{|c|c|c|c|c|}\hline 1 & 2 & 3 & 4 & \underset{\uparrow}{10} \\ \hline \end{array}$$

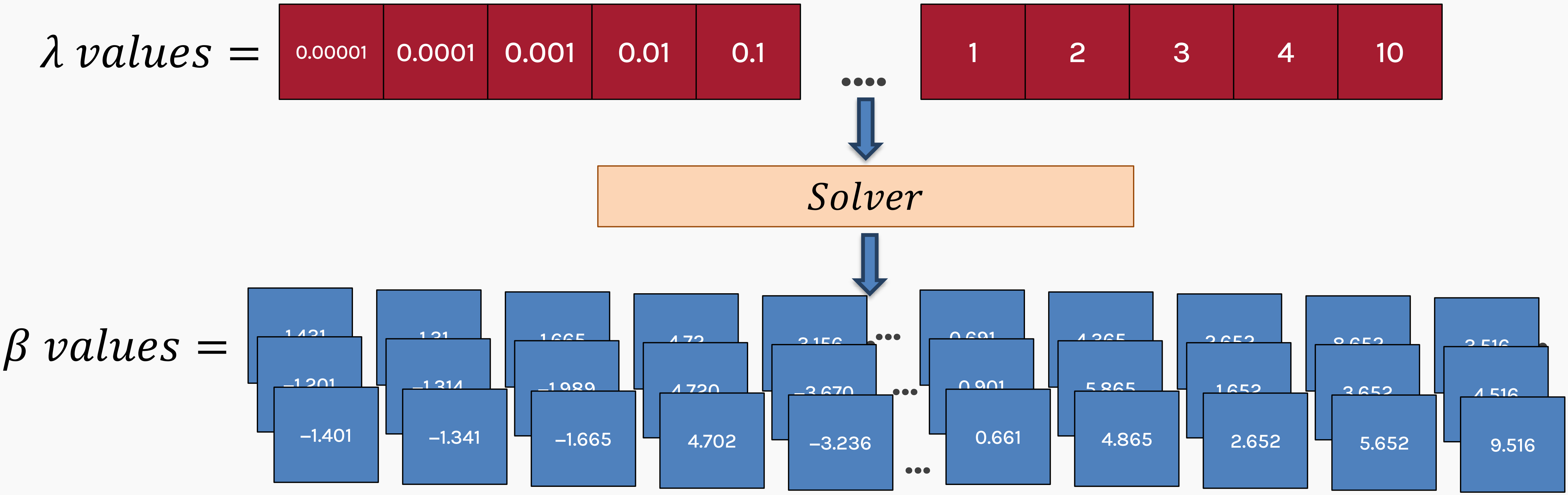
$$\lambda_{min} \qquad\qquad\qquad\qquad\qquad\qquad \lambda_{max}$$

Lasso regularization with only **validation** : step by step

Step 3 : Determine β

Train

For each λ value, we determine $\beta_{Lasso}(\lambda)$ using train data



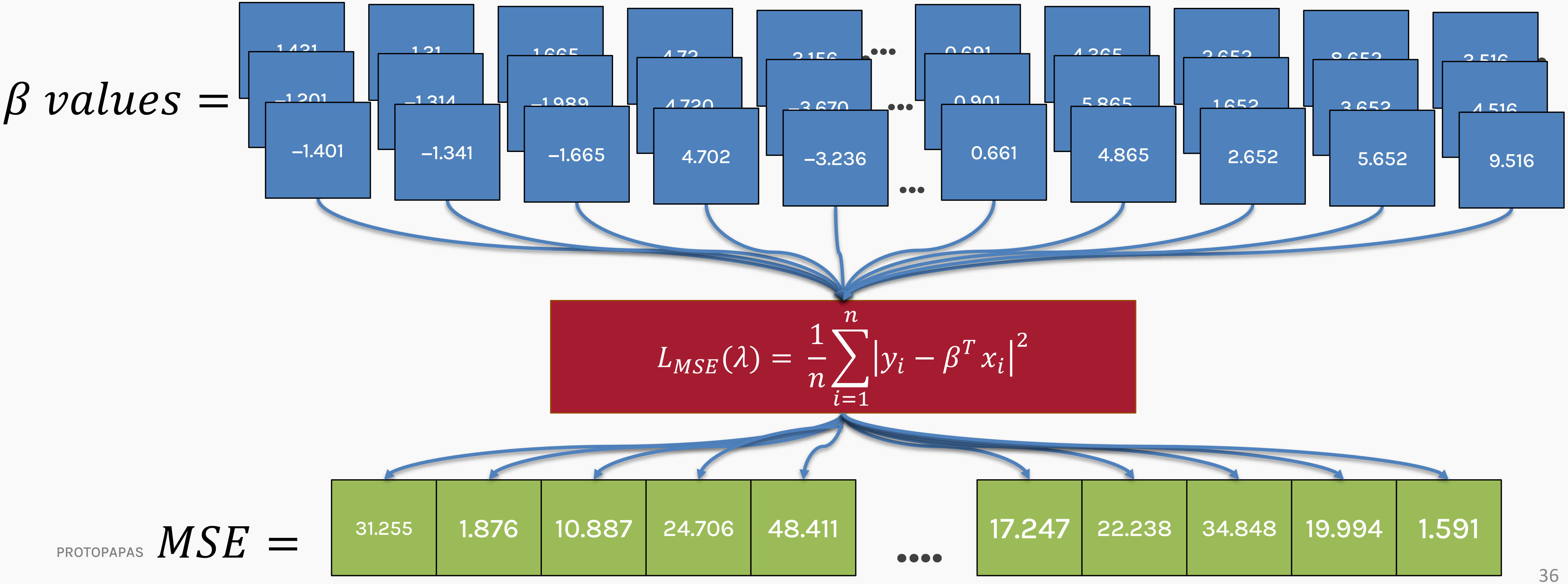
Keep in mind that each λ value gives us a different vector of β coefficients.

Lasso regularization with only **validation** : step by step

Validation

Step 4 : Record the $L_{MSE}(\lambda)$

For each vector of β coefficients, we record $L_{MSE}(\lambda)$ using the validation data



Lasso regularization with only **validation** : step by step

Step 5 : Select the λ_{Lasso}

Validation

Select the λ that minimizes the MSE loss on the validation data.

$\lambda\ values =$

0.00001	0.0001	0.001	0.01	0.1	....	1	2	3	4	10

$MSE =$

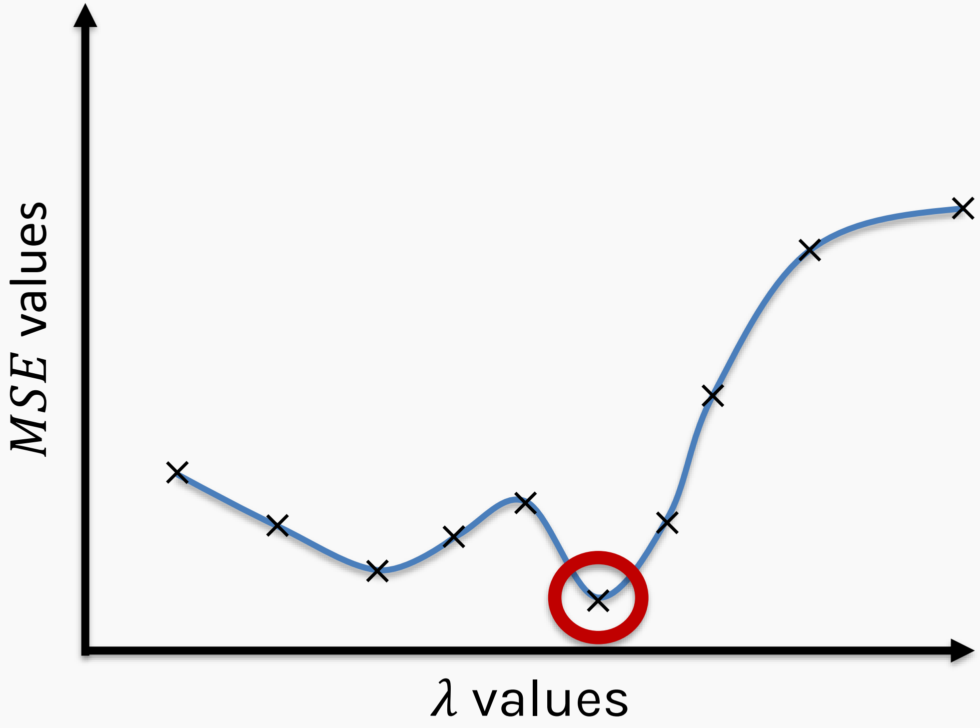
31.255	1.876	10.887	24.706	48.411	....	17.247	22.238	34.848	19.994	1.591

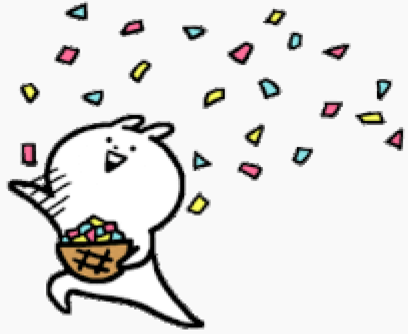
Let's visualize and see it!

Step 5 : Select the λ_{Lasso}

Validation

Select the λ that minimizes the MSE loss on the validation data.

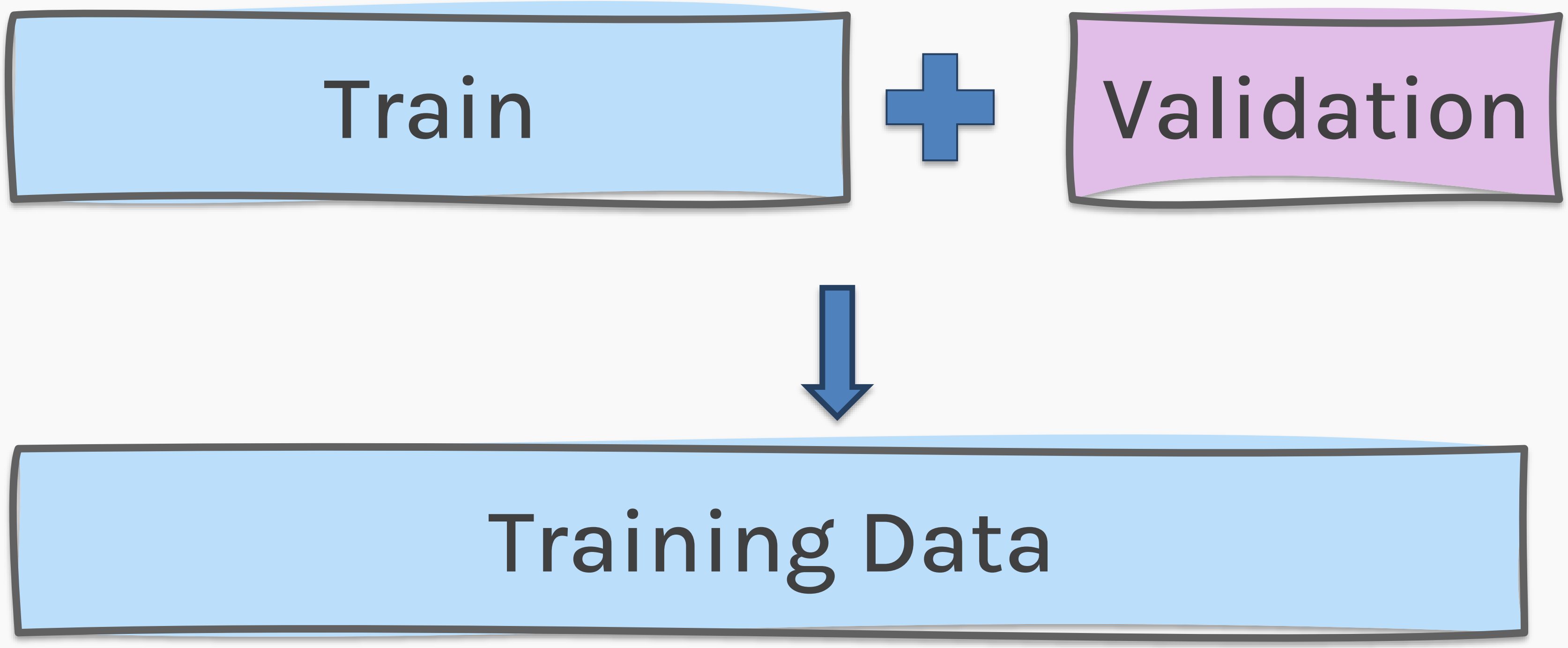




Lasso regularization with only **validation** : step by step

Step 6 : Refit the model

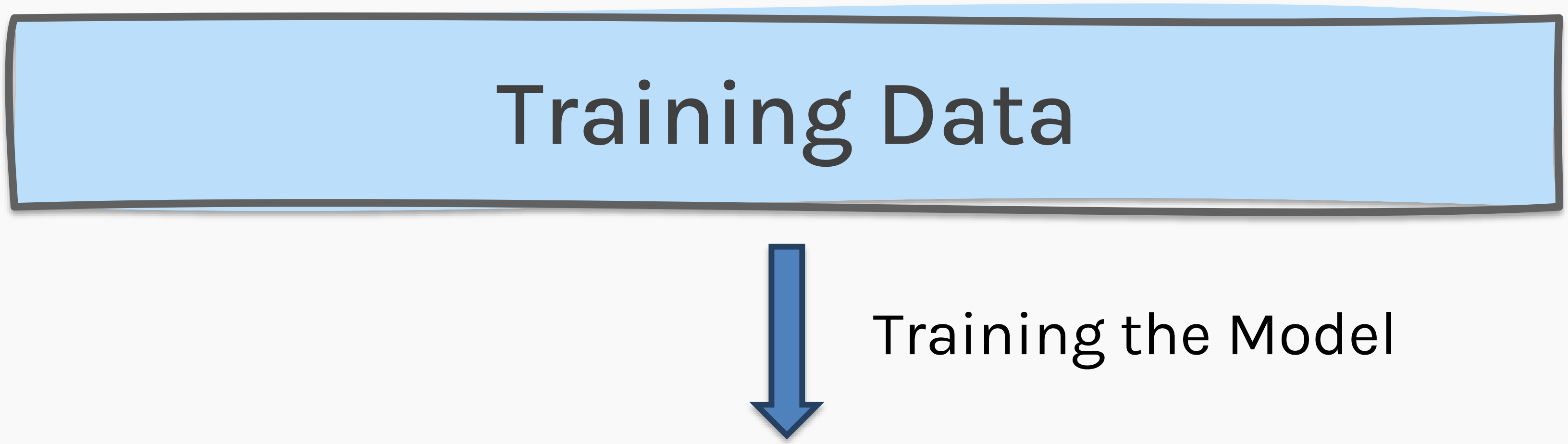
Refit the model using both train and validation data using λ_{Lasso}.



Step 6 : Refit the model

Refit the model using both train and validation data using λ_{Lasso}.

This gives us $\widehat{\beta}_{Lasso}(\lambda)$



$$\widehat{\beta}_{Lasso}(\lambda)$$

Lasso regularization with only **validation** : step by step

Step 7 : Record MSE/R2

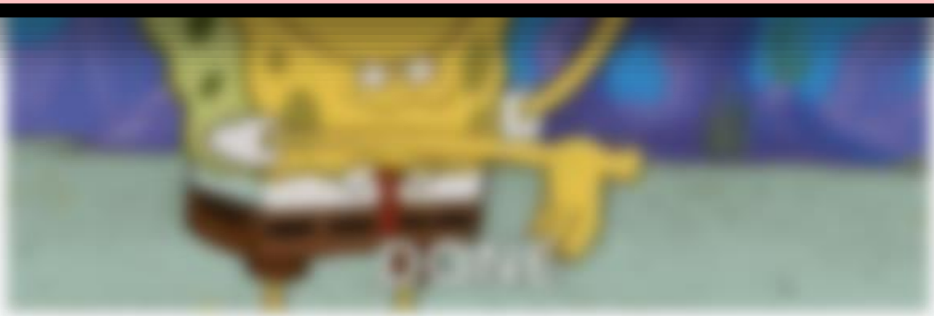
Report MSE or R2 on test data given the $\widehat{\boldsymbol{\beta}}_{\boldsymbol{Lasso}}(\boldsymbol{\lambda})$

Lasso regularization with only validation : step by step

Step 7 : Record MSE/R2

Report MSE or R2 on test data given the $\hat{\beta}_{Lasso}(\lambda)$

Instead of relying on a single validation set, we can use cross-validation to get a more reliable estimate of the best λ



Ridge regularization with **CV**: step by step

Regularization with **CV**: step by step

Step 1 : Split Training Data

Split training data into K folds.

Regularization with **CV**: step by step

Step 2: Select a fold as Validation

Select one fold as validation and the rest as training data

Train	Train	Train	Train	Valid.

Step 3: Select a range of possible λ values

$\lambda\ values =$

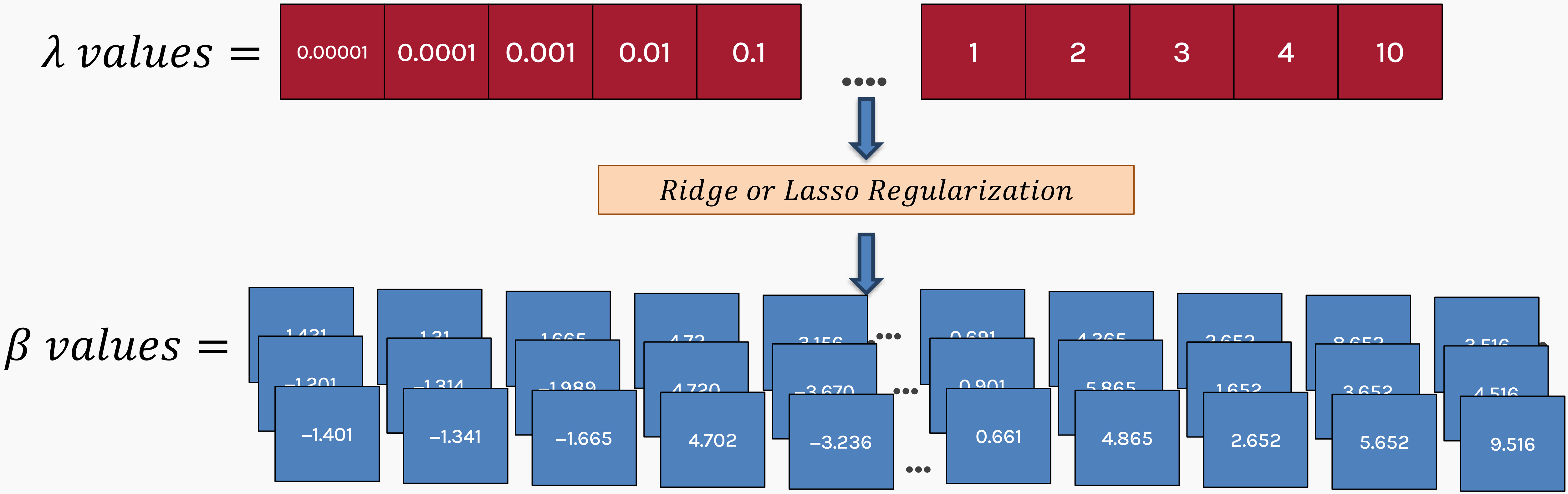
0.00001	0.0001	0.001	0.01	0.1	....	1	2	3	4	10

λ_{min} λ_{max}

Step 3 : Determine β

Train

For each λ value, we determine $\beta_{Lasso/Ridge}(\lambda)$ using train data



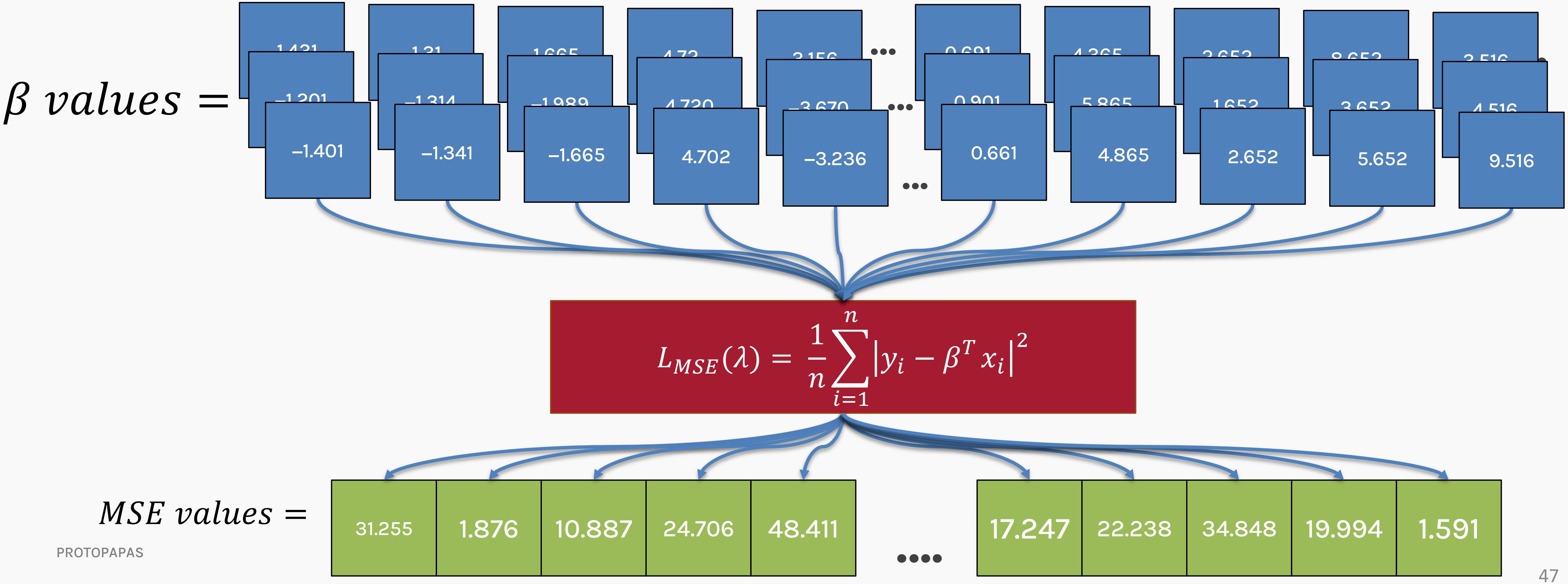
Keep in mind that each λ value gives us a different vector of β coefficients.

Regularization with **CV**: step by step

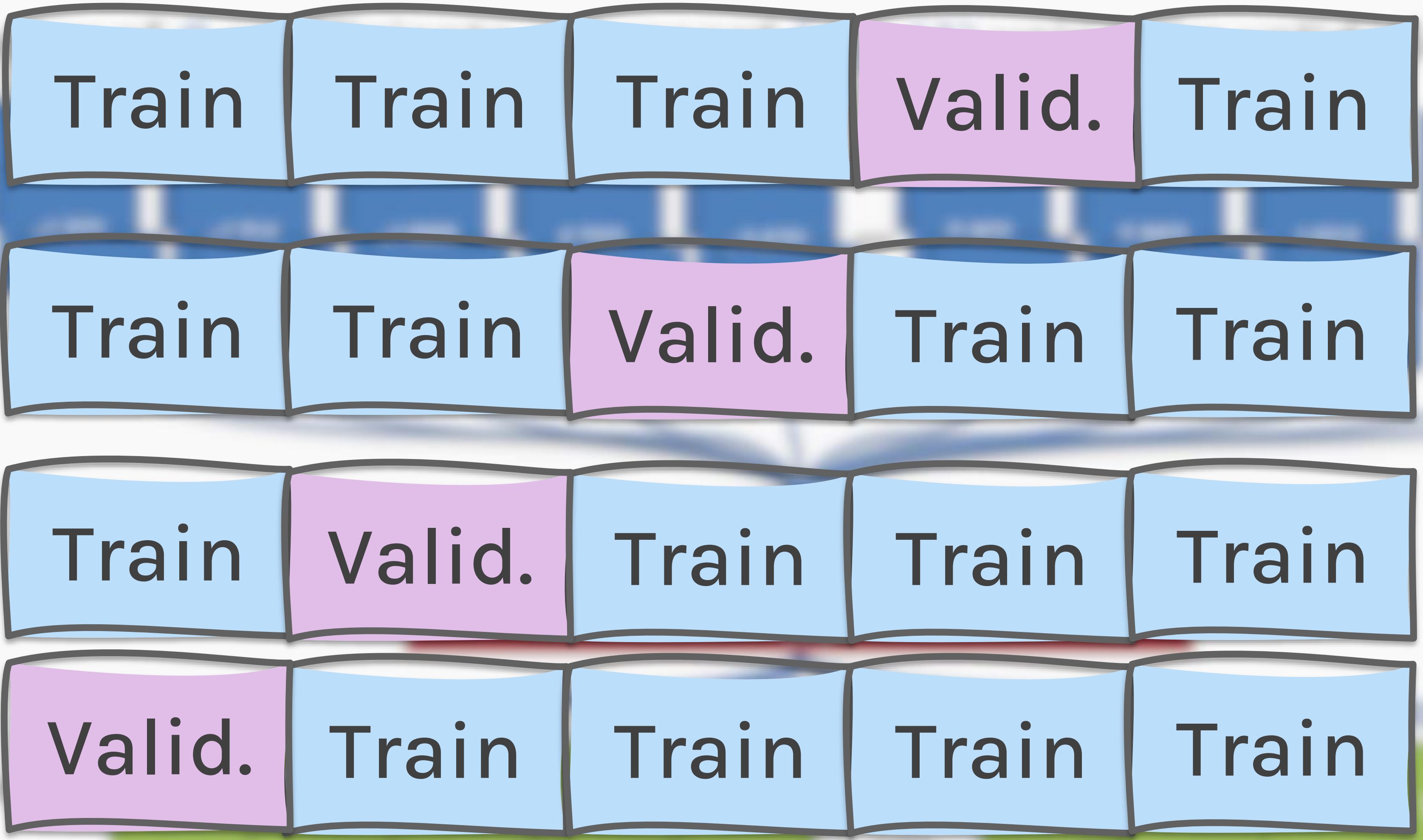
Validation

Step 4 : Record the $L_{MSE}(\lambda)$

For each vector of β coefficients, we record $L_{MSE}(\lambda)$ using the validation data



Repeat Step 2 to 4 for different values of k



Regularization with **CV**: step by step

Step 6: Average MSE, $\bar{L}_{MSE}(\lambda)$

Calculate the average MSE, $\bar{L}_{MSE}(\lambda)$ for each λ by averaging $L_{MSE}(\lambda, k)$ over k. folds.

$MSE_1 =$	31.255	1.876	10.887	24.706	48.411	....	17.247	22.238	34.848	19.994	1.591
$MSE_2 =$	33.255	1.676	6.887	4.706	38.411	....	7.247	2.238	3.848	1.994	3.591
⋮						⋮					
$MSE_k =$	36.255	3.876	4.887	5.706	8.411	....	1.247	5.238	4.848	9.994	4.591
$Average\ MSE =$	33.521	2.476	7.550	11.706	31.744	....	8.58	9.904	14.515	10.660	3.257

Regularization with **CV**: step by step

Step 7: Select λ

Find the λ that minimizes the $\bar{L}_{MSE}(\lambda)$

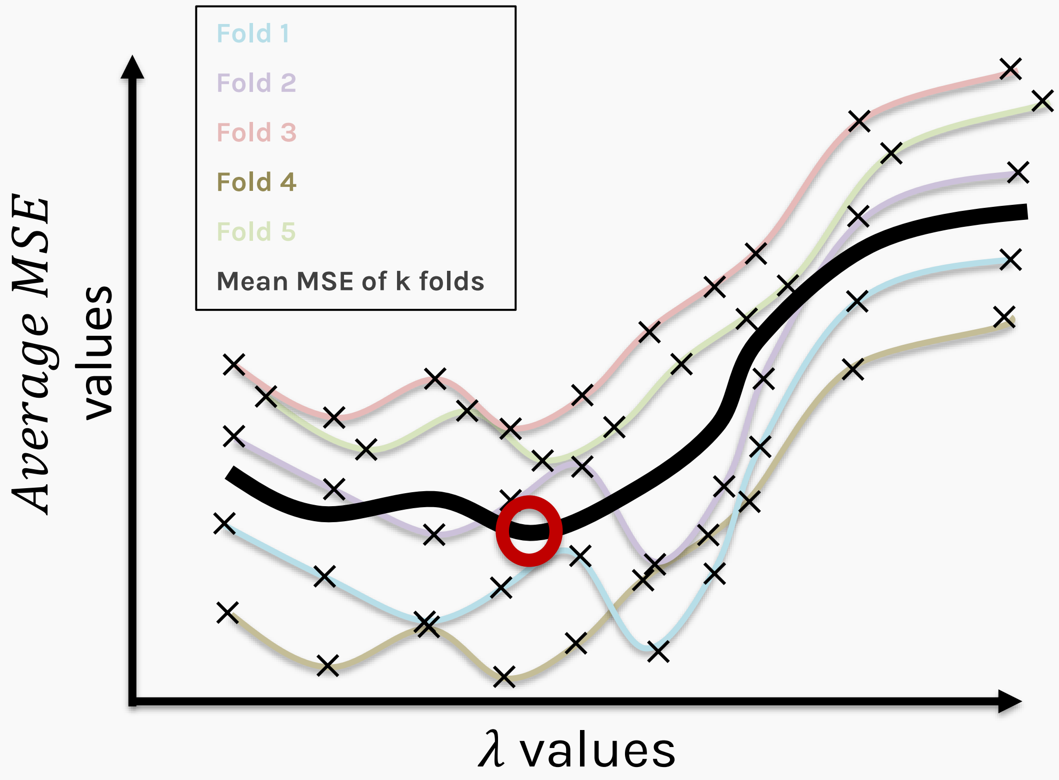
$Average\ MSE =$	33.521	2.476	7.550	11.706	31.744	....	8.58	9.904	14.515	10.660	3.257
$\lambda\ values =$	0.00001	0.0001	0.001	0.01	0.1	....	1	2	3	4	10

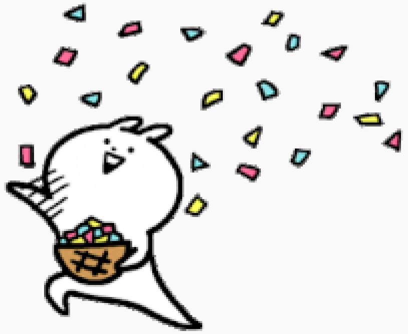
Let's visualize and see it!

Regularization with **CV**: step by step

Step 7: Select λ

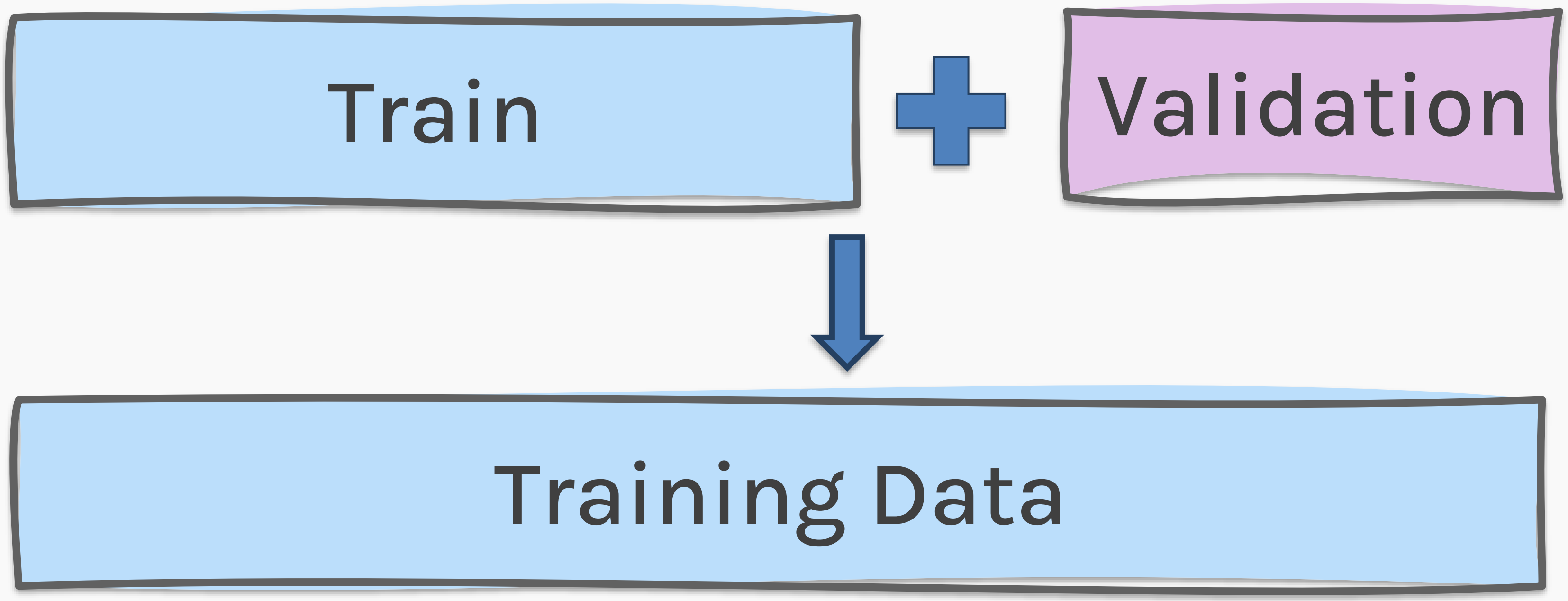
Find the λ that minimizes the $\bar{L}_{MSE}(\lambda)$





Step 8: Refit the model

Refit the model using both train and validation data using λ.



Regularization with **CV**: step by step

Step 8: Refit the model

Refit the model using both train and validation data using λ.

This gives us $\widehat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$$

Regularization with **CV**: step by step

Step 9 : Record MSE/R2

Report MSE or R2 on test data given the $\widehat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$

Ridge regularization with only **validation** : step by step

For ridge regression there exist an analytical solution for the coefficients:

$$\hat{\beta}_{Ridge}(\lambda) = \left(\mathrm{X}^{\mathrm{T}}\mathrm{X} + \lambda I\right)^{-1} X^T Y$$

1. split data into $\{\{X,Y\}_{train}, \{X,Y\}_{validation}, \{X,Y\}_{test}\}$
2. for λ in $\{\lambda_{min}, \dots \lambda_{max}\}$:
    1. determine the β that minimizes the L_{ridge}, $\beta_{Ridge}(\lambda) = \left(\mathrm{X}^{\mathrm{T}}\mathrm{X} + \lambda I\right)^{-1} X^T Y$, using the train data.
    2. record $L_{MSE}(\lambda)$ using validation data.
3. select the λ that minimizes the *MSE* loss on the validation data,

$$\lambda_{ridge} = \mathrm{argmin}_{\lambda}\, L_{MSE}(\lambda)$$

4. Refit the model using both **train and validation data**, $\{\{X,Y\}_{train}, \{X,Y\}_{validation}\}$, **now using** λ_{ridge}, **resulting to** $\hat{\beta}_{ridge}(\lambda_{ridge})$
5. Report MSE or R^2 on $\{X,Y\}_{test}$ given the $\hat{\beta}_{ridge}(\lambda_{ridge})$

Lasso regularization with **validation** only: step by step

For Lasso regression, there is **no** analytical solution for the coefficients, so we use a **solver**.

1. split data into $\{\{X,Y\}_{train}, \{X,Y\}_{validation}, \{X,Y\}_{test}\}$
2. for λ in $\{\lambda_{min}, \ldots \lambda_{max}\}$:
   A. determine the β that minimizes the L_{lasso}, $\beta_{lasso}(\lambda)$, using the train data. **This is done using a solver.**
   B. record $L_{MSE}(\lambda)$ using the validation data.
3. select the λ that minimizes the ***MSE* loss** on the validation data,

$$\lambda_{lasso} = \mathrm{argmin}_{\lambda} L_{MSE}(\lambda)$$

4. Refit the model using both train and validation data, $\{\{X,Y\}_{train}, \{X,Y\}_{validation}\}$, now using λ_{Lasso}, resulting to $\hat{\beta}_{lasso}(\lambda_{lasso})$
5. Report MSE or R^2 on $\{X,Y\}_{test}$ given the $\hat{\beta}_{lasso}(\lambda_{lasso})$

Ridge regularization with CV: step by step

	λ_1	λ_2	…	λ_n
k_1	L_{11}	L_{12}	..	…
k_2	L_{21}	…	..	…
…	..	…	..	…
k_n	…	…	…	…
E[]	$\bar{L}_1$	$\bar{L}_2$	…	$\bar{L}_n$

1. remove $\{X,Y\}_{test}$ from data
2. split the rest of data into K folds, $\{\{X,Y\}_{train}^{-k}, \{X,Y\}_{val}^{k}\}$
3. for k in $\{1,\dots,K\}$

   for λ in $\{\lambda_0, \dots, \lambda_n\}$:

   A. determine the β that minimizes the L_{ridge}, $\beta_{ridge}(\lambda,k) = \left(X^T X + \lambda I\right)^{-1} X^T Y$, using the train data of the fold, $\{X,Y\}_{train}^{-k}$.

   B. record $L_{MSE}(\lambda,k)$ using the validation data of the fold $\{X,Y\}_{val}^{k}$

At this point we have a 2-D matrix, rows are for different k, and columns are for different λ values.

4. Calculate the average MSE, $\bar{L}_{MSE}(\lambda)$ the for each λ by averaging $L_{MSE}(\lambda,k)$ over k folds.
5. Find the λ that minimizes the $\bar{L}_{MSE}(\lambda)$, resulting to λ_{ridge}.
6. Refit the model using the full training data, $\{\{X,Y\}_{train}, \{X,Y\}_{val}\}$, resulting to $\hat{\beta}_{ridge}(\lambda_{ridge})$
7. report MSE or R^2 on $\{X,Y\}_{test}$ given the $\hat{\beta}_{ridge}(\lambda_{ridge})$