Model Selection with Cross Validation

CS109A Introduction to Data Science

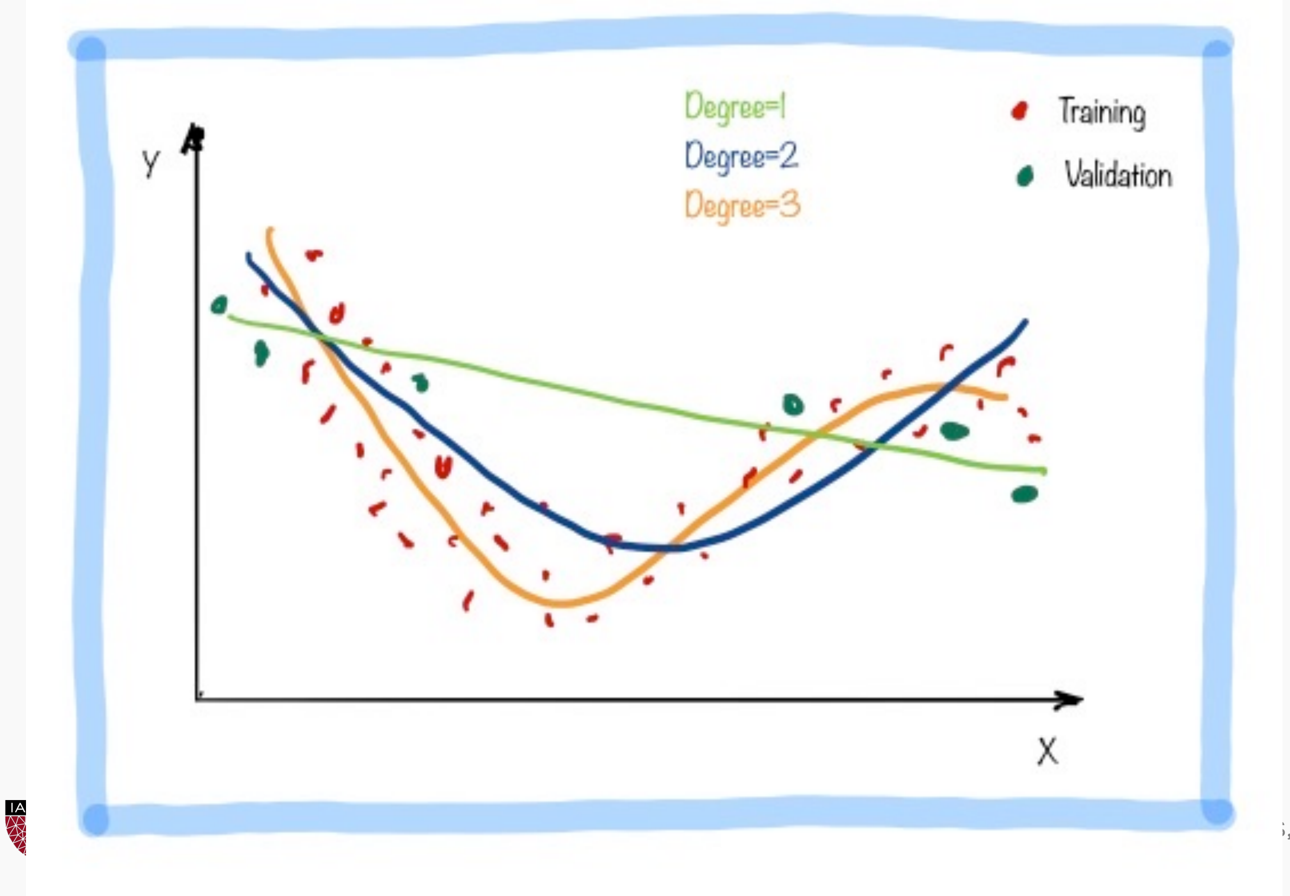
Pavlos Protopapas, Natesh Pillai

COUNTY STATS FOR TEXAS

County	7-day average cases ↑	7-day average deaths	Cases	Deaths
Glasscock County	-1	0	134	3
Borden County	0	0	34	2
Goliad County	0	0	568	19

Cross Validation: Motivation

Using a single validation set to select amongst multiple models can be problematic - **there is the possibility of overfitting to the validation set**.



It is obvious that degree=3 is the correct model but the validation set by chance favors the linear model.

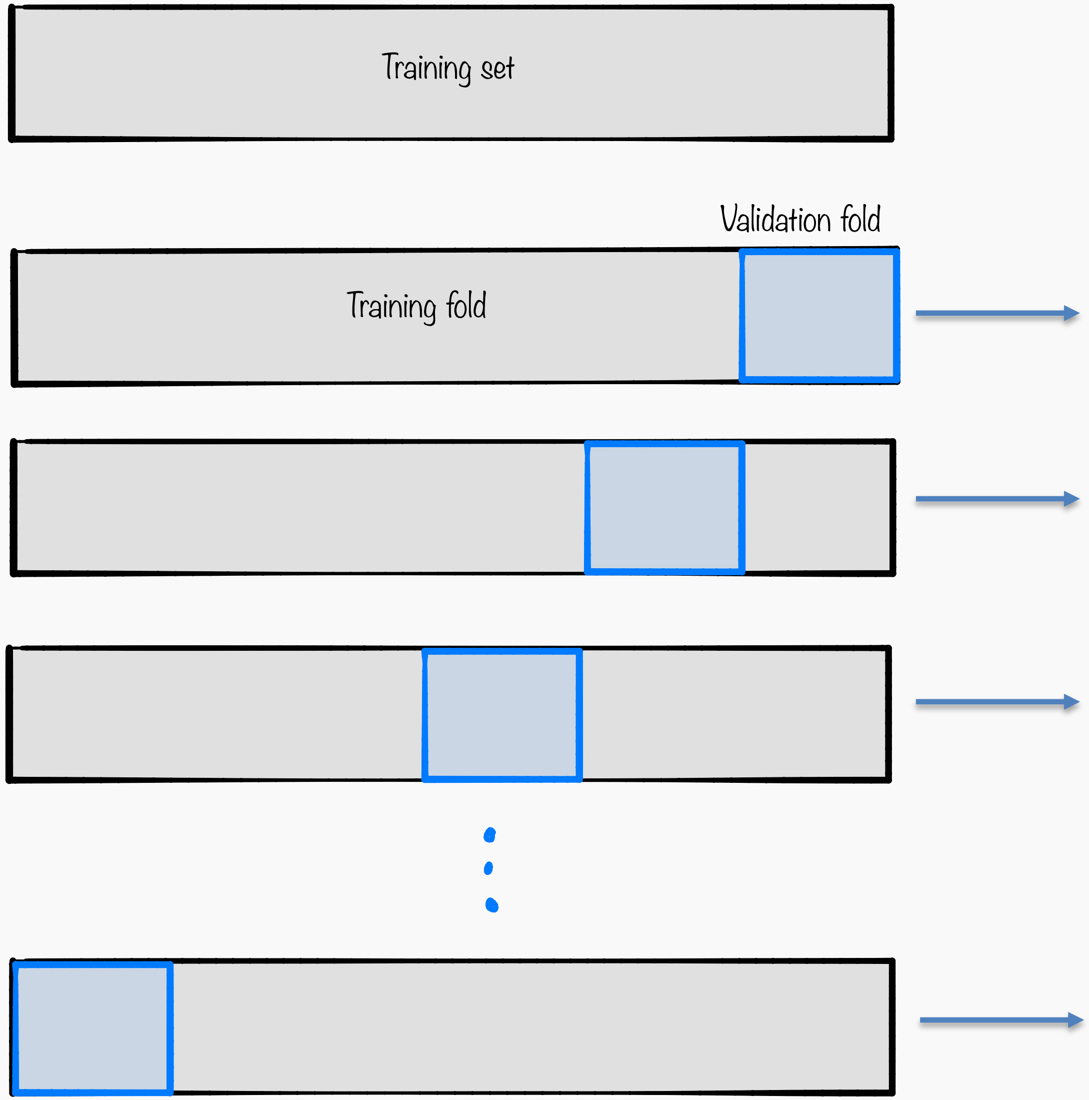
Cross Validation: Motivation

Using a single validation set to select amongst multiple models can be problematic - **there is the possibility of overfitting to the validation set**.

One solution to the problems raised by using a single validation set is to evaluate each model on **multiple** validation sets and average the validation performance.

One can randomly split the training set into training and validation multiple times **but** randomly creating these sets can create the scenario where important features of the data never appear in our random draws.

Cross Validation



$$\mathrm{MSE}_1^{\mathrm{val}}$$

$$\mathrm{MSE}_2^{\mathrm{val}}$$

$$\mathrm{MSE}_3^{\mathrm{val}}$$

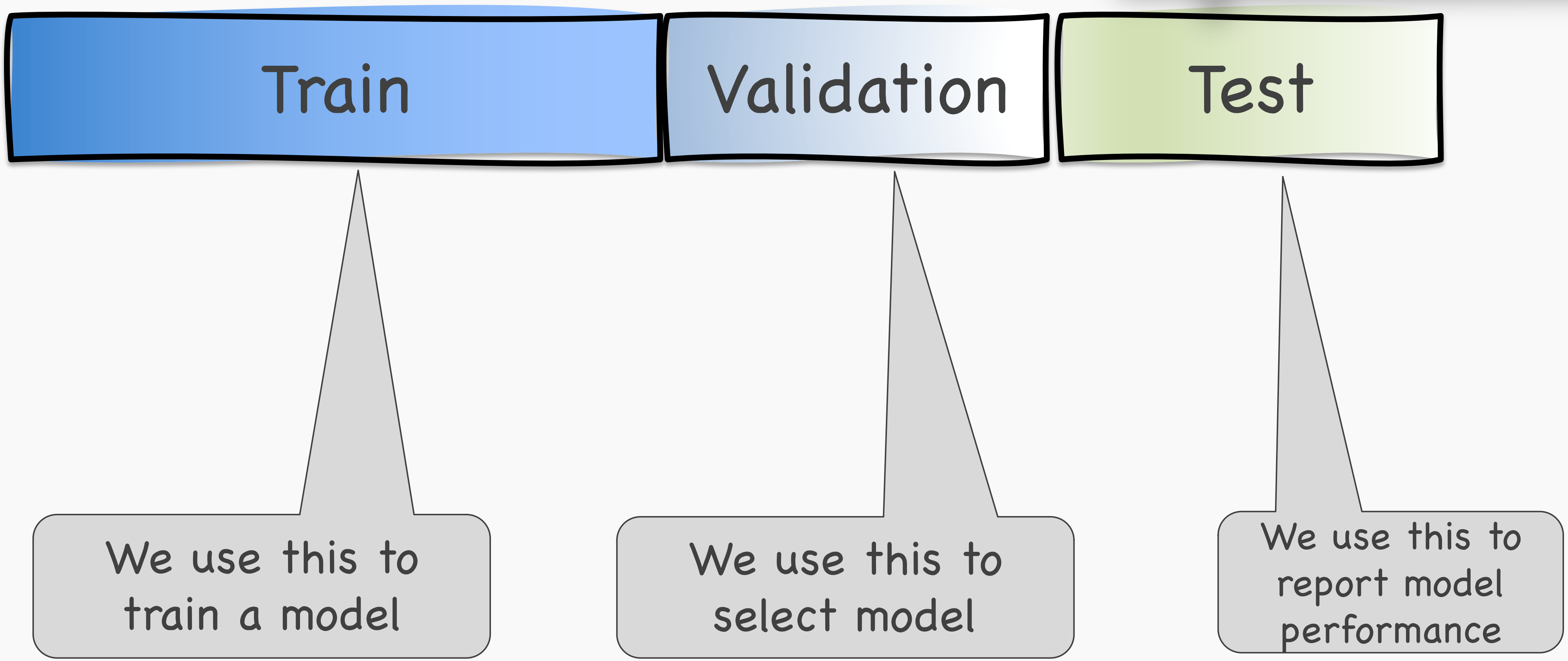
$$\mathrm{MSE}_k^{\mathrm{val}}$$

$$\mathrm{MSE}^{\mathrm{val}} = \frac{1}{k}\sum_{i=1}^{k} \mathrm{MSE}_i^{\mathrm{val}}$$

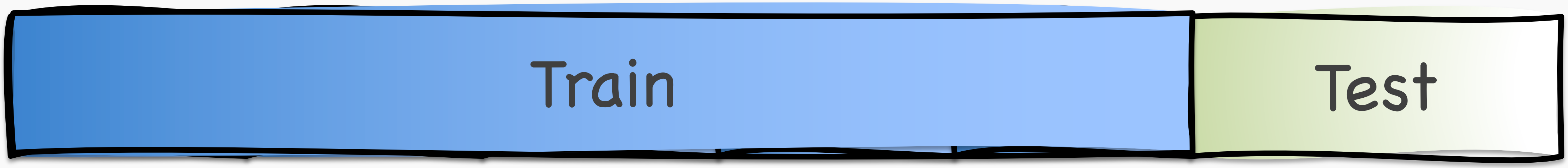
Train-Validation-Test

We introduce a different sub-set, which we called validat
to select the model.

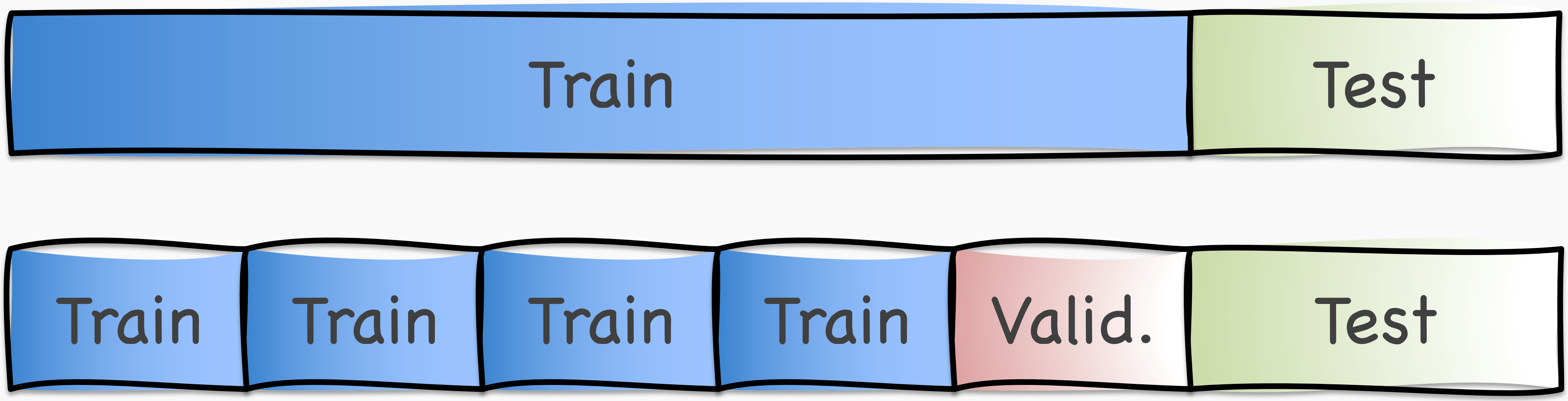
The test set should never be touched for model training or selection.



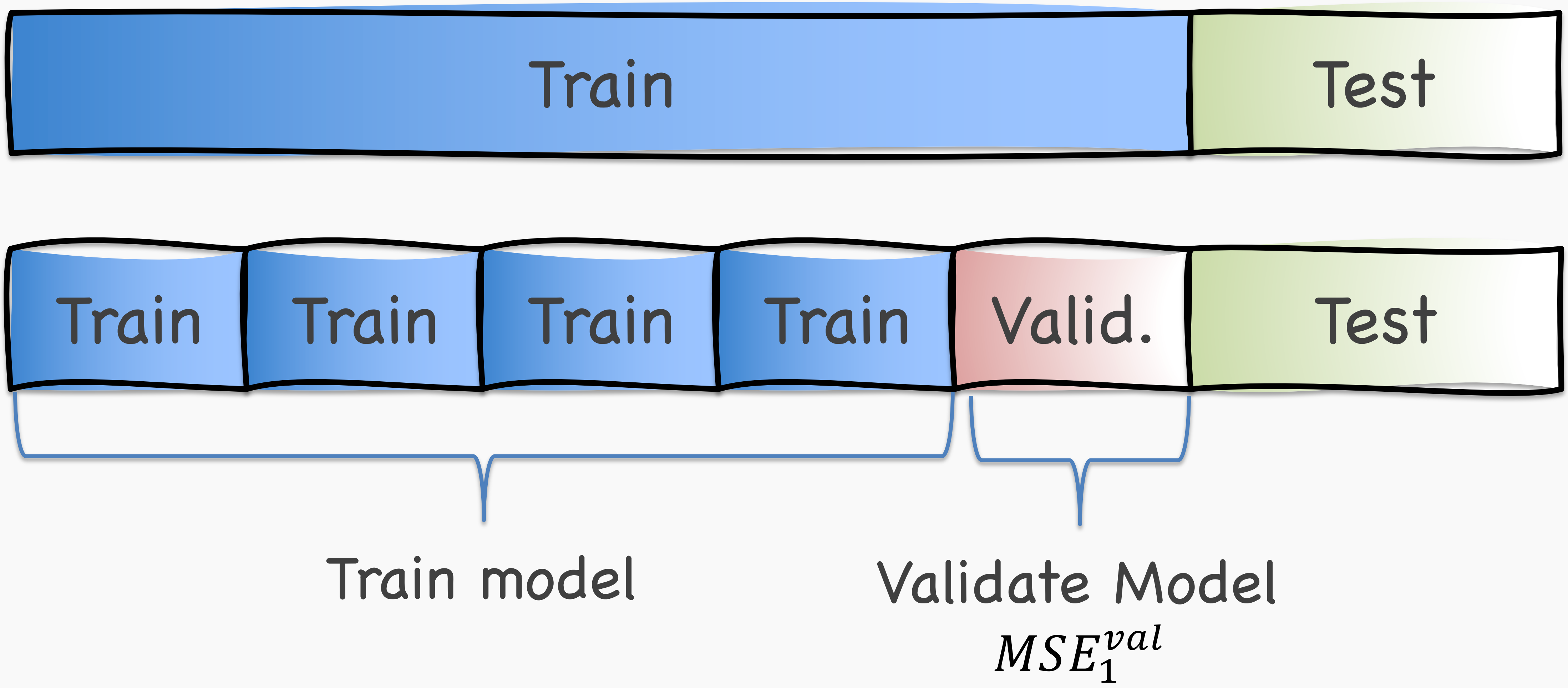
Cross Validation



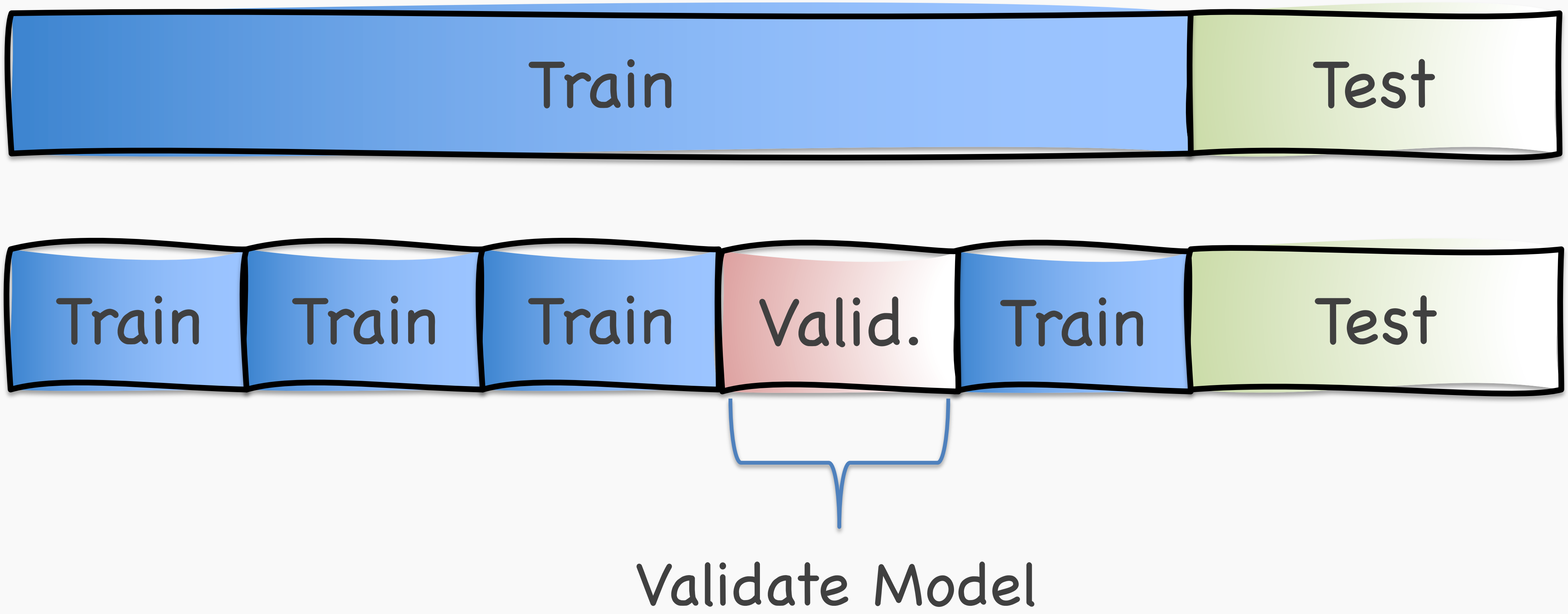
Cross Validation



Cross Validation

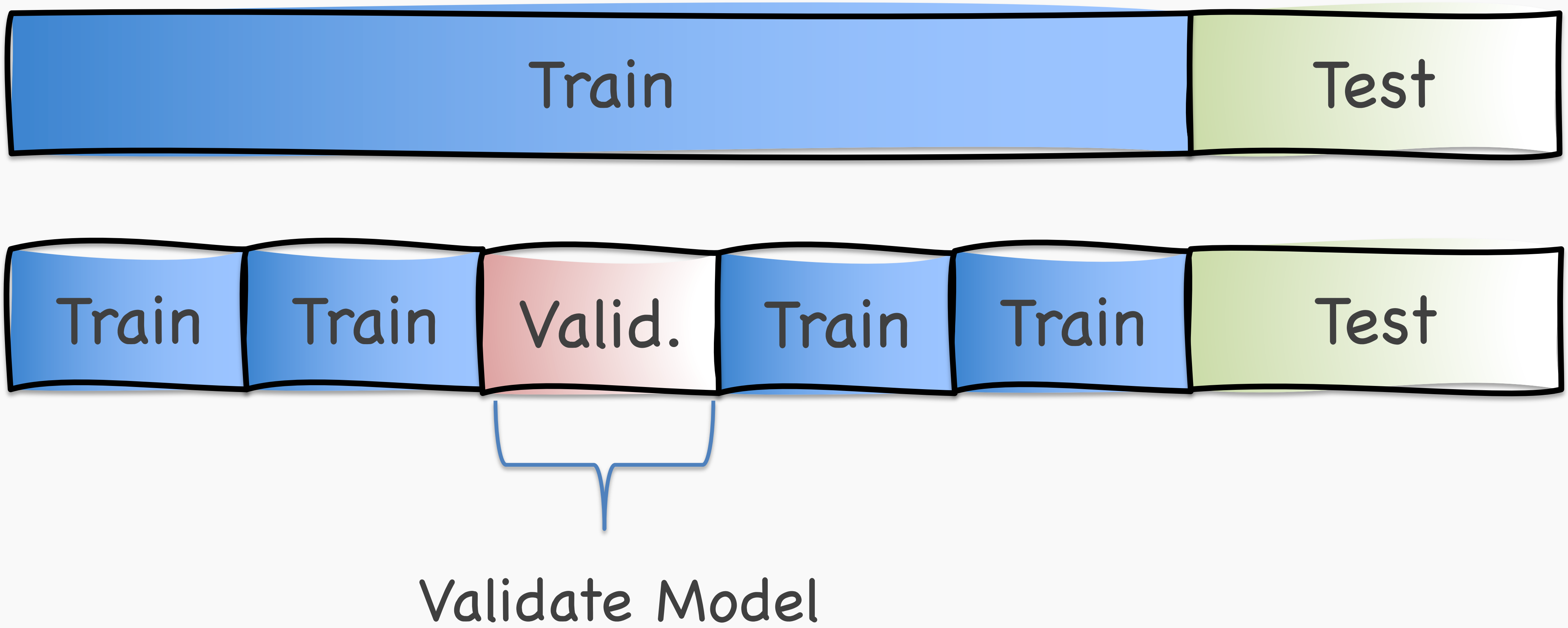


Cross Validation



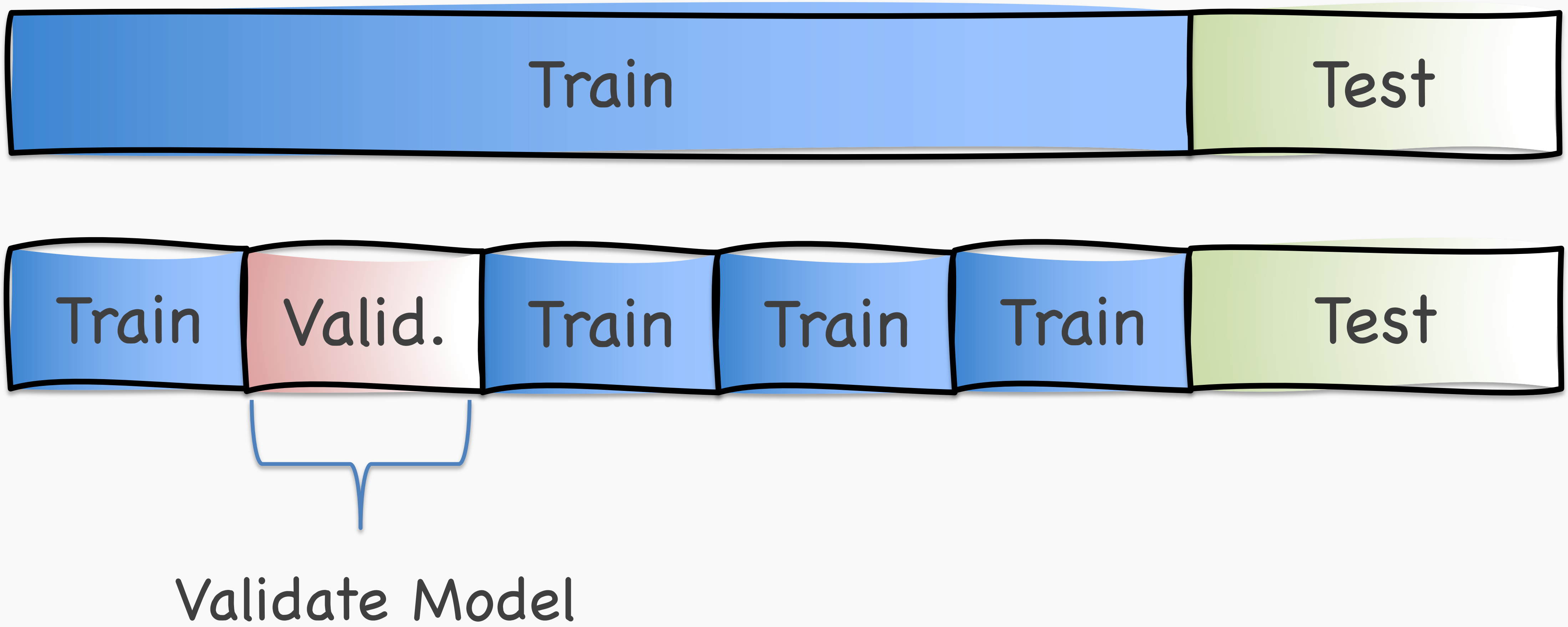
$$MSE_2^{val}$$

Cross Validation



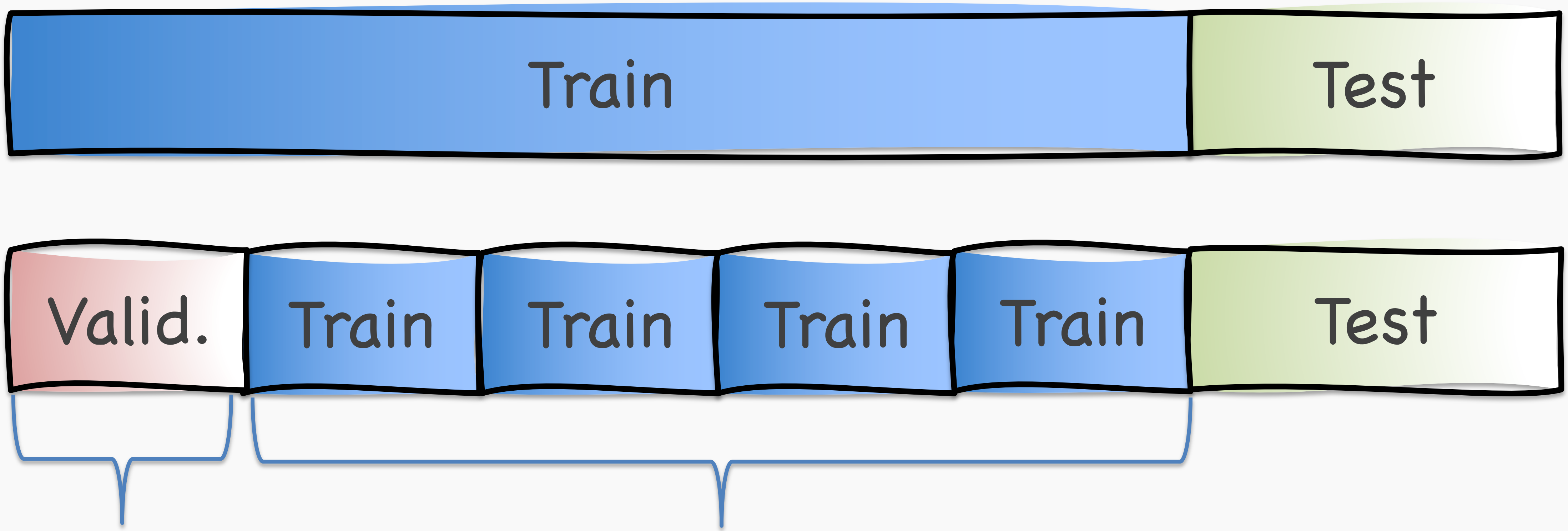
$$MSE_3^{val}$$

Cross Validation



$$MSE_4^{val}$$

Cross Validation



Validate Model

$$MSE_5^{val}$$

Train model

$$MSE^{val} = \frac{1}{5}\sum_{i=1}^{5} MSE_i^{val}$$

K-Fold Cross Validation

Given a data set $\{X_1, \ldots, X_n\}$, where each $\{X_1, \ldots, X_n\}$ contains J features.

To ensure that every observation in the dataset is included in at least one training set and at least one validation set we use the **K-fold validation**:

- split the data into K uniformly sized chunks, $\{C_1, \ldots, C_K\}$
- we create K number of training/validation splits, using one of the K chunks for validation and the rest for training.

We fit the model on each training set, denoted $\hat{f}_{C_{-i}}$, and evaluate it on the corresponding validation set, $\hat{f}_{C_{-i}}(C_i)$. The ***cross validation is the performance*** of the model averaged across all validation sets:

$$CV(\text{Model}) = \frac{1}{K}\sum_{i=1}^{K} L(\hat{f}_{C_{-i}}(C_i))$$

where L is a loss function.

Leave-One-Out

Or using the ***leave one out*** method:

- validation set: $\{X_i\}$
- training set: $X_{-i} = \{X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n\}$

for $i = 1, \dots, n$:

We fit the model on each training set, denoted $\hat{f}_{X_{-i}}$, and evaluate it on the corresponding validation set, $\hat{f}_{X_{-i}}(X_i)$.

The ***cross validation score*** is the performance of the model averaged across all validation sets:

$$CV(\text{Model}) = \frac{1}{n}\sum_{i=1}^{n} L(\hat{f}_{X_{-i}}(X_i))$$

where L is a loss function.

The model used to fit the data

The data to fit

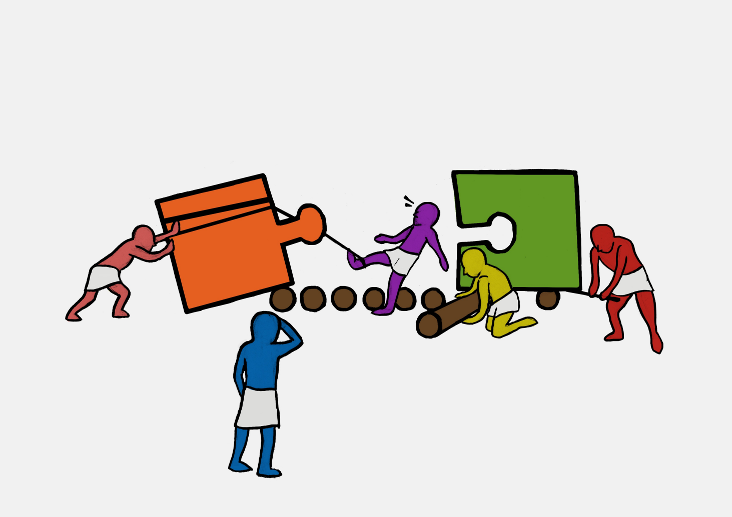
The target variable to predict on

```
sklearn.model_selection.cross_validate(estimator, X, y,
    scoring, cv, return_train_score)
```

Number of folds

Strategy to evaluate the performance of the cross-validated model on the test set.
Use "neg_mean_squared_error" for regression

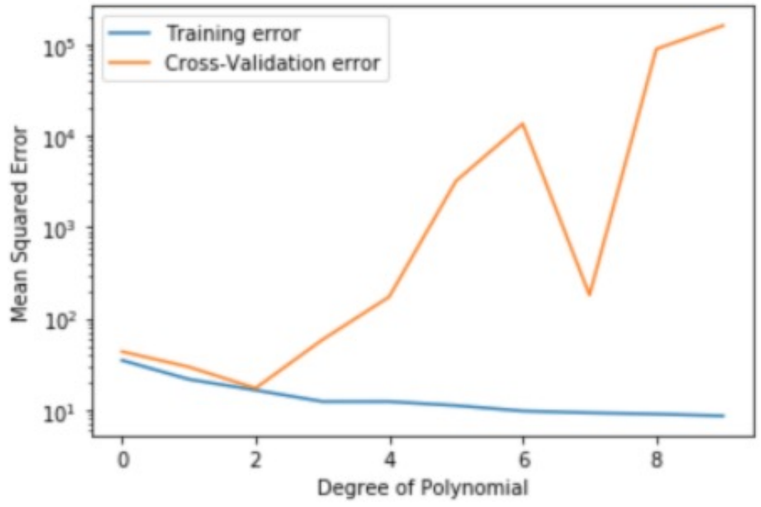
Set to True to include train scores



Description

Exercise: Best Degree of Polynomial using Cross-validation

The aim of this exercise is to find the **best degree** of polynomial based on the MSE values. Further, plot the train and cross-validation error graphs as shown below.



Instructions:

- Read the dataset and split into train and validation sets.
- Select a max degree value for the polynomial model.
- For each degree:
  - Perform k-fold cross validation
  - Fit a polynomial regression model for each degree on the training data and predict on the validation data
- Compute the train, validation and cross-validation error as MSE values and

When to use CV and when to use Validation only?

Choosing number of folds?

Scaling: Revisited