

## 8: Resampling Methods (K-fold Cross Validation)

```
$ echo "Data Science Institute"
```

# Activity

Watch this two videos about  $k$ -Fold Cross-Validation:

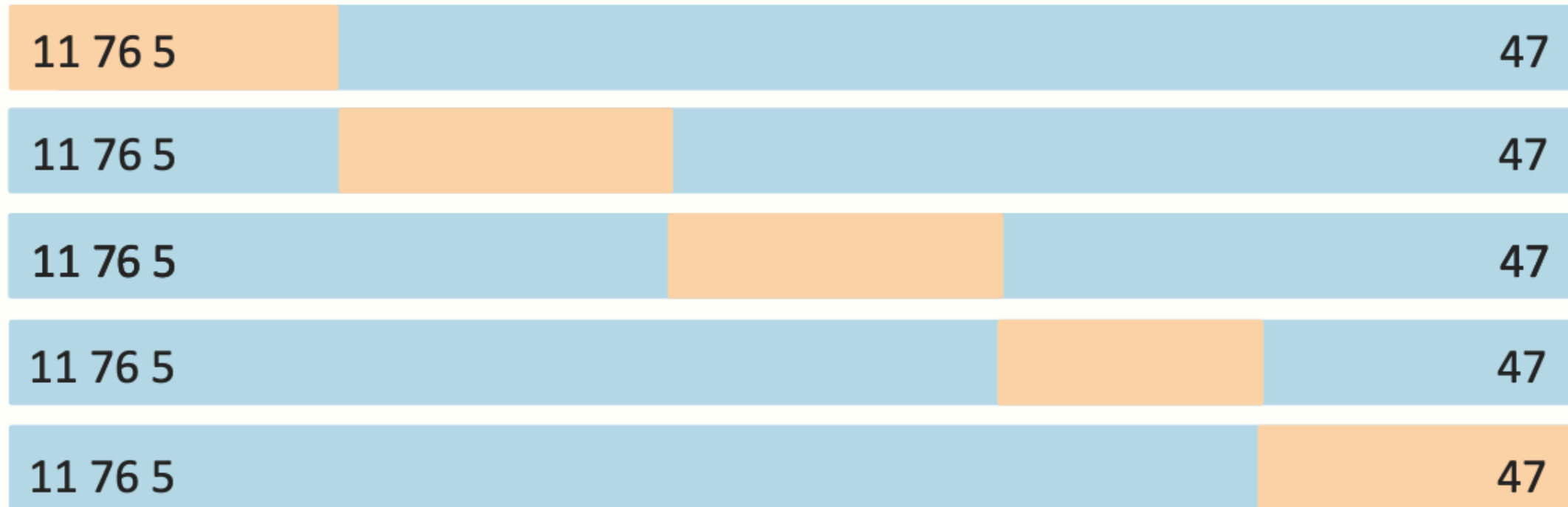
- <https://www.youtube.com/watch?v=hoNpvry0370> (~3 minutes)
- <https://www.youtube.com/watch?v=wjILv3-UGM8&t=439s> (~10 minutes)

# $k$ -Fold Cross-Validation

$k$ -fold cross validation involves randomly dividing the set of observations into  $k$  *approximately* equally sized groups. Then,

- Fit the model using the observations from all but one of the groups.
- Make predictions for the response of the observations in the remaining group.
- Compute the validation set error.
- Repeat this process for each group.

# $k$ -Fold Cross-Validation



# $k$ -Fold Cross-Validation

The  $k$ -fold cross validation estimate of the test error is the average of the  $k$  validation set errors.

$$CV_{(k)} = \frac{1}{k} \sum_{i=1}^k MSE_i$$

LOOCV is a special case of the  $k$ -fold cross validation approach using  $k = n$  where  $n$  is the number of observations.

## Breakout Room

Based on what you learned so far, how does LOOCV compare to  $k$ -fold CV?

# Comparing LOOCV and $k$ -fold CV

- The ♦ *computational time/effort for  $k$ -fold CV for  $k < n$  is less* ♦ since we are fitting fewer models in the process.
- ♦ *LOOCV is less biased* ♦ in its estimation of the test error rate since it trains the model on more observations.

# Comparing LOOCV and $k$ -fold CV Continue

- ♦ LOOCV has a test error estimate that has higher variance ♦ than  $k$ -fold CV ( $k < n$ )
  - The models in the LOOCV process are fitted with *nearly* identical training sets thus each test error result is much more correlated with one another than they would be for  $k$ -fold CV
  - Averaging highly correlated quantities has a higher variance than if they were not correlated

Thus, there is a bias-variance trade-off when it comes to choosing  $k$  for  $k$ -fold cross-validation. Typically  $k = 5$  or  $k = 10$  is used.



## Exercises: $k$ -fold CV

Open the  $k$ -fold CV Jupyter Notebook file.

- Go over the " $k$ -fold CV" section together as a class.
- Questions should be completed at home if time does not allow.

# References

Chapter 5 of the ISLP book:

James, Gareth, et al. "Resampling Methods." An Introduction to Statistical Learning: with Applications in Python, Springer, 2023.