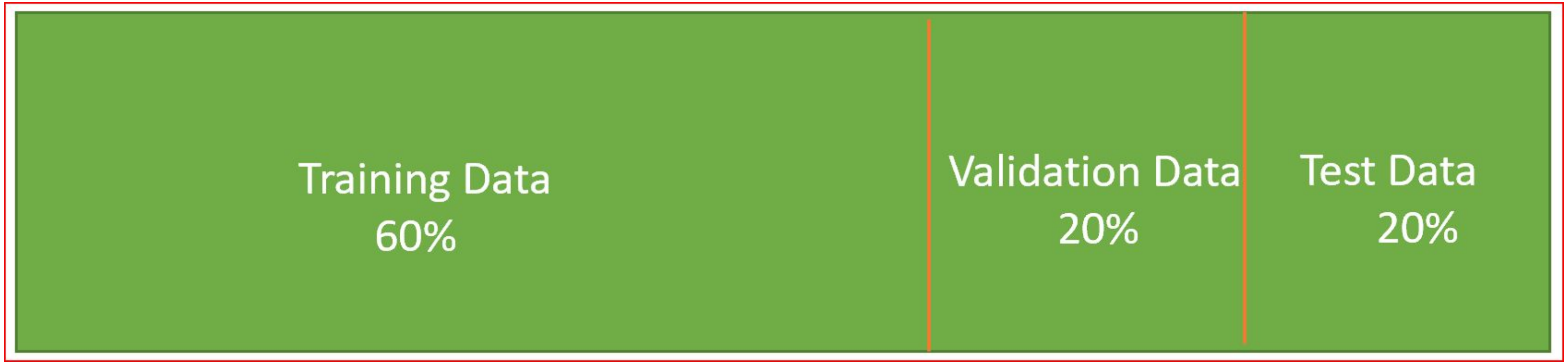
Cross Validation

Presentation by Berk Sudan

Remember: a typical split



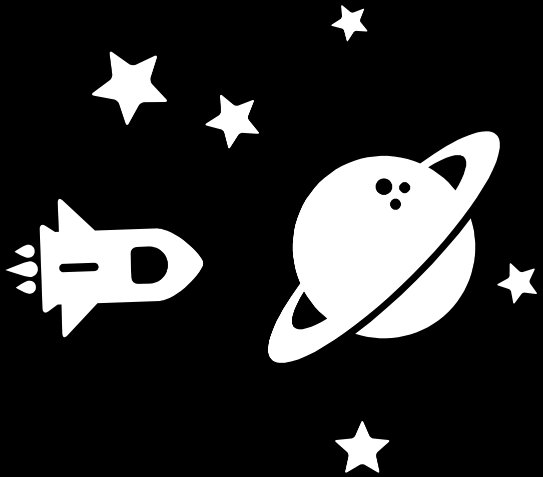
Training Data
60%

Validation Data
20%

Test Data
20%

Ref: *https://medium.com/datadriveninvestor/k-fold-and-other-cross-validation-techniques-6c03a2563f1e*

Why we use cross validation?



Fail to Generalize (Overfitting)

Due to ***sample variability between training and test set***:

- Better prediction on **training data** but fail to **generalize** on **test data**.
- So, **low training error rate** but **a high test error rate**.

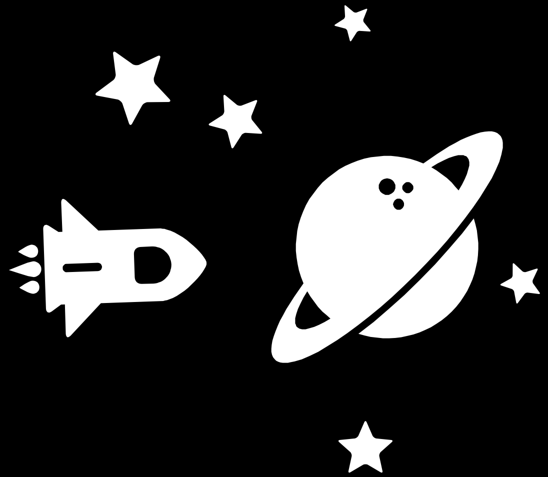
Ref: *https://datanee.com/2016/08/11/outlier-detection-an-overview-and-applications/*

Overestimation of Test Set

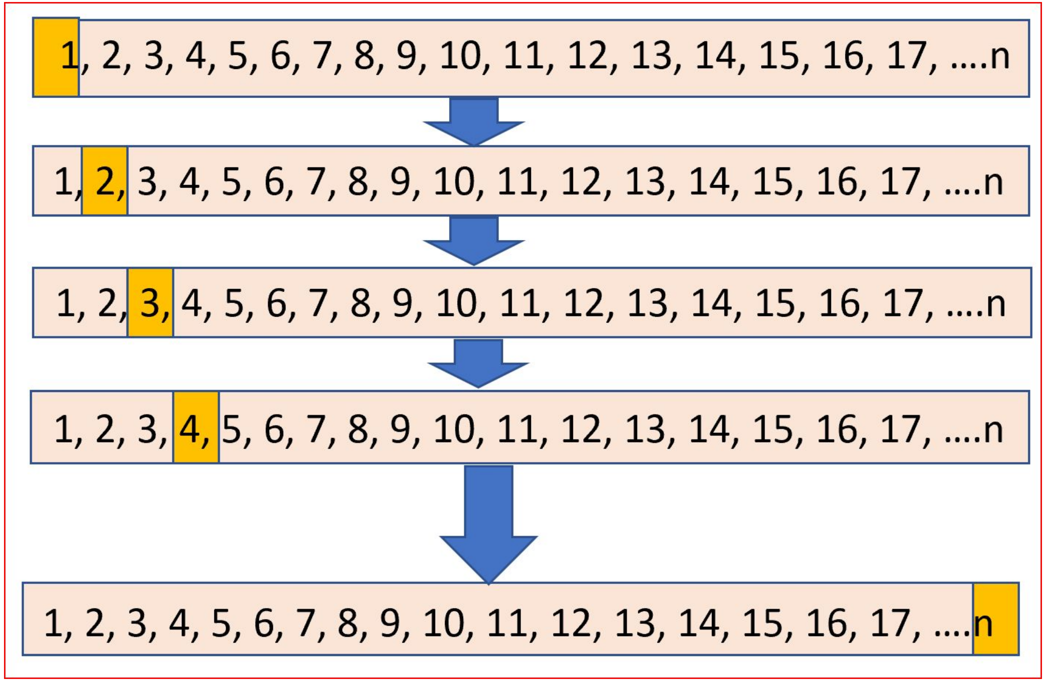
- We use only a **subset** of data, i.e **fewer observations**
- So, **overestimates** the test error rate!

Ref: *https://datanee.com/2016/08/11/outlier-detection-an-overview-and-applications/*

Cross Validation Methods



Leave One out Cross Validation — LOOCV



1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, ….n

1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, ….n

1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, ….n

1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, ….n

1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, ….n

Ref: *https://towardsdatascience.com/a-brief-overview-of-outlier-detection-techniques-1e0b2c19e561*

Leave One out Cross Validation — LOOCV

+. **Far less bias** as we have used the **entire** dataset

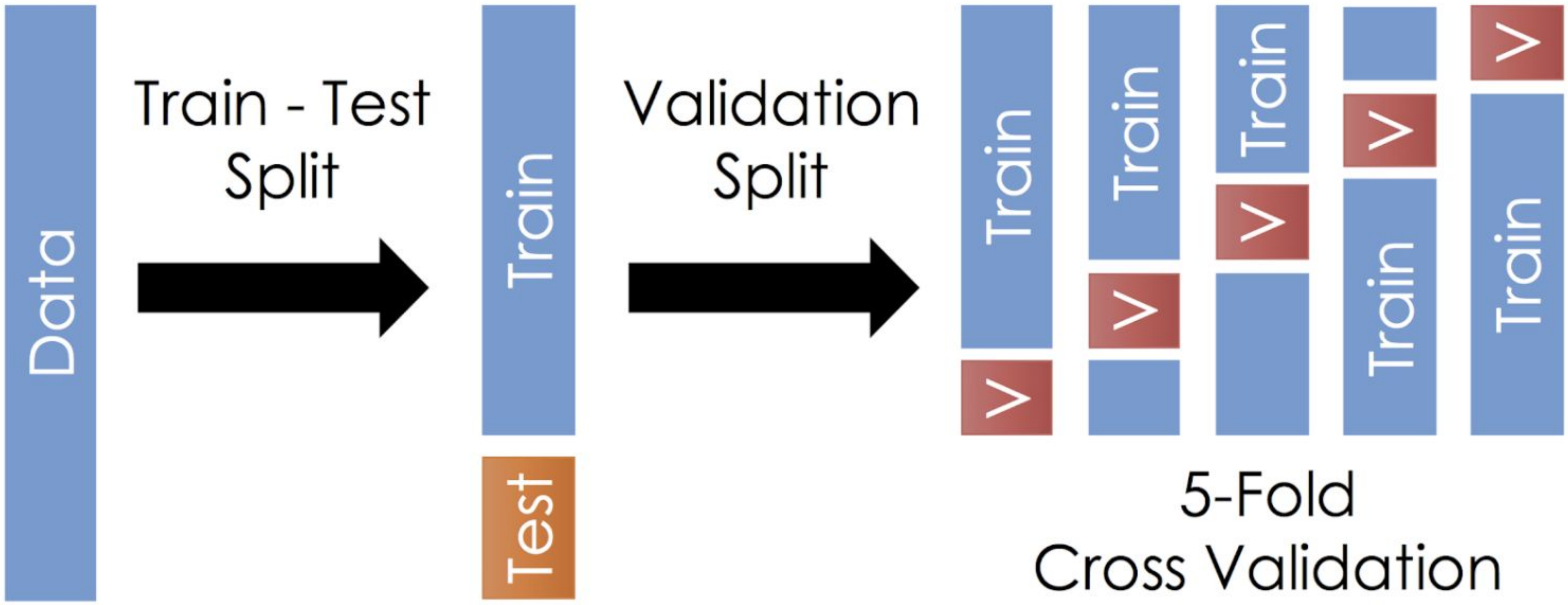
+. **No randomness** in the training/test data

- . MSE will vary, if test point is **outlier**!

- . Execution is expensive -> **O(n) complexity**

Ref: *https://towardsdatascience.com/a-brief-overview-of-outlier-detection-techniques-1e0b2c19e561*

K-Fold Cross Validation (e.g 5-fold)



K-Fold Cross Validation

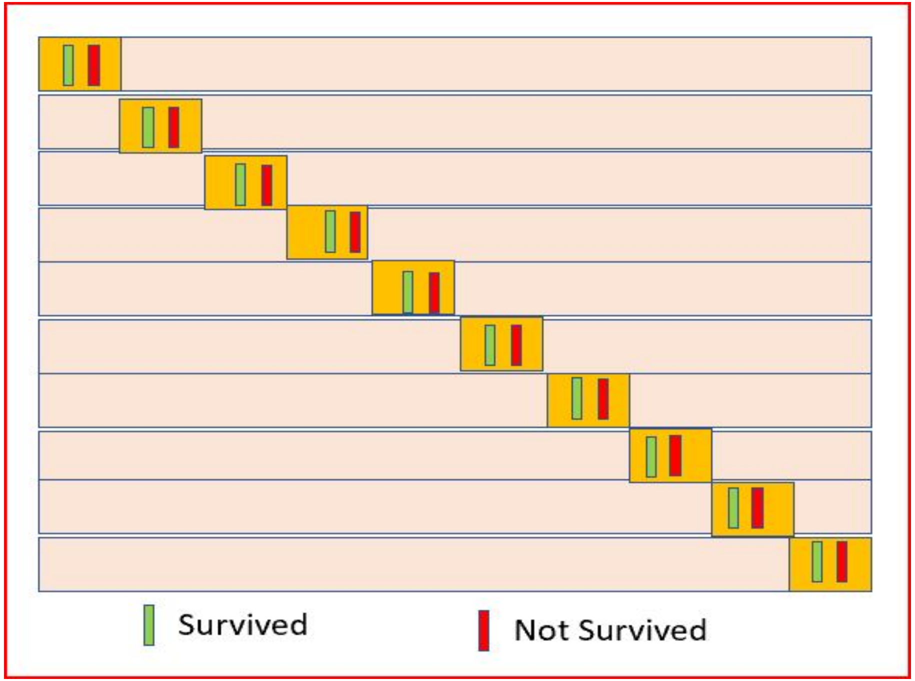
+. **Computation time** is **reduced**

+. Reduced **bias**

- . In **unbalanced** datasets, test-set may not **represent** the dataset

Ref: *https://towardsdatascience.com/a-brief-overview-of-outlier-detection-techniques-1e0b2c19e561*

Stratified K-Fold Cross Validation



Ref: *https://towardsdatascience.com/a-brief-overview-of-outlier-detection-techniques-1e0b2c19e561*

Stratified K-Fold Cross Validation

+. **Helps** with **reducing** both **bias** and **variance**

- . **More** computation

Ref: *https://towardsdatascience.com/a-brief-overview-of-outlier-detection-techniques-1e0b2c19e561*

End of Presentation

Presented by Berk Sudan