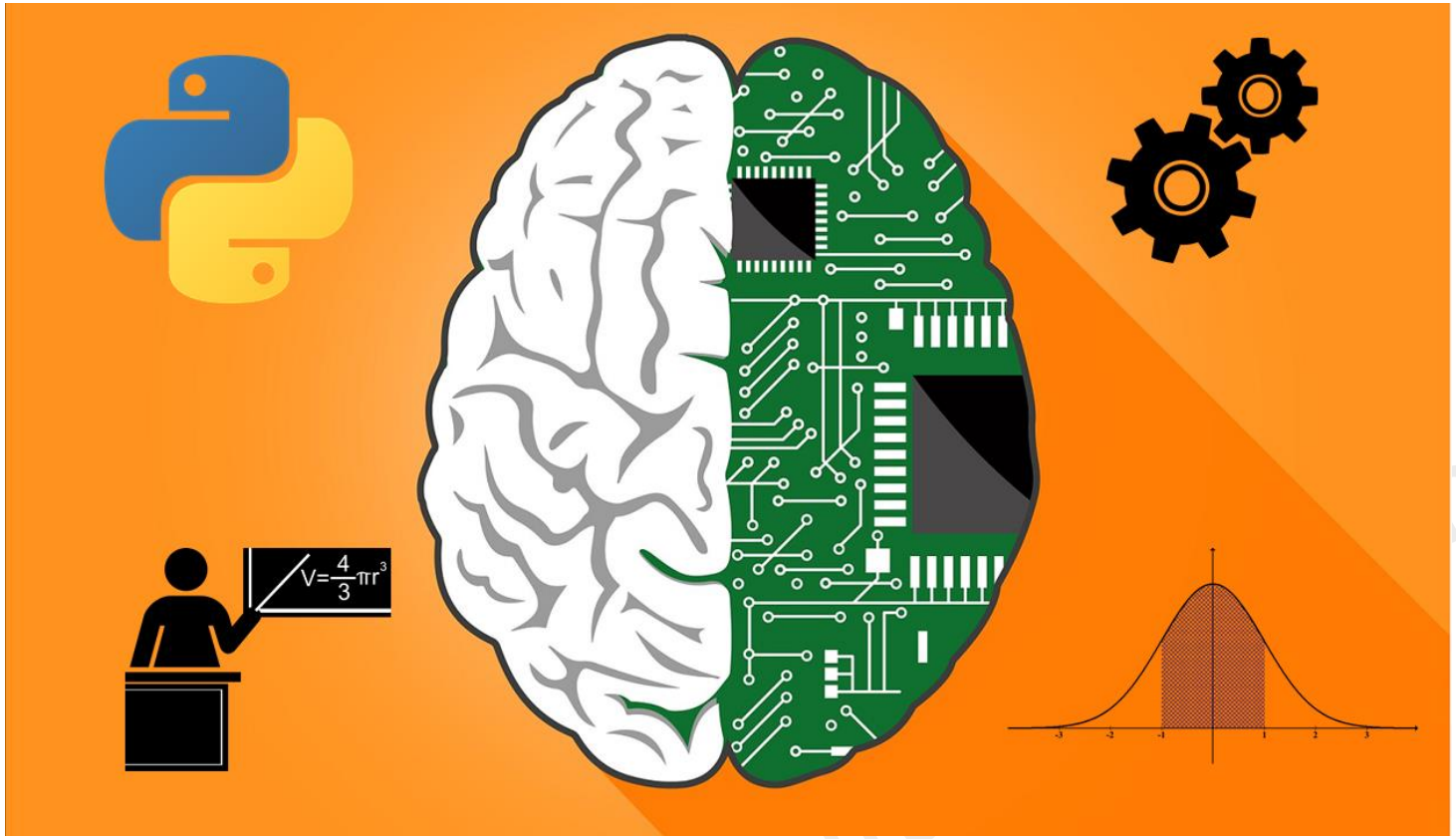
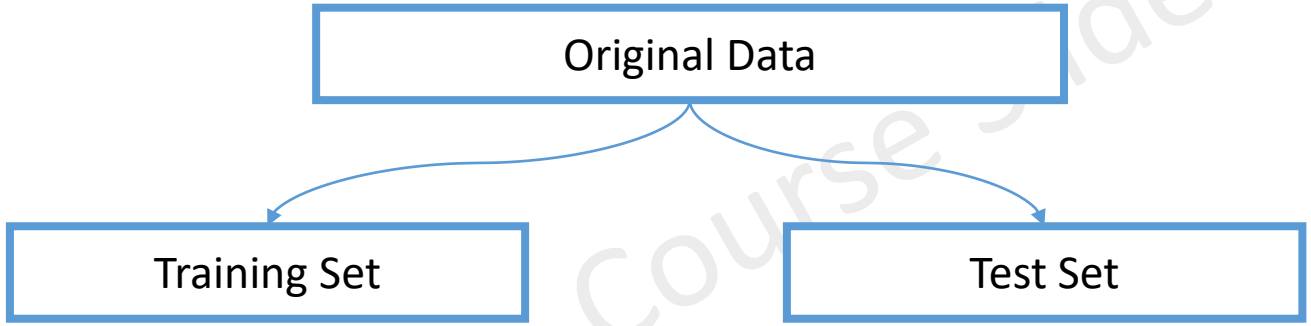


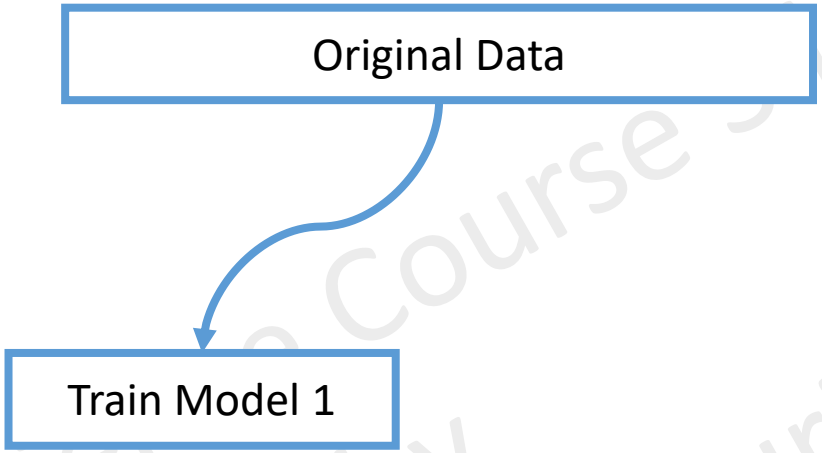
Complete Data Science and Machine Learning Using Python

By
Jitesh Khurkhuriya

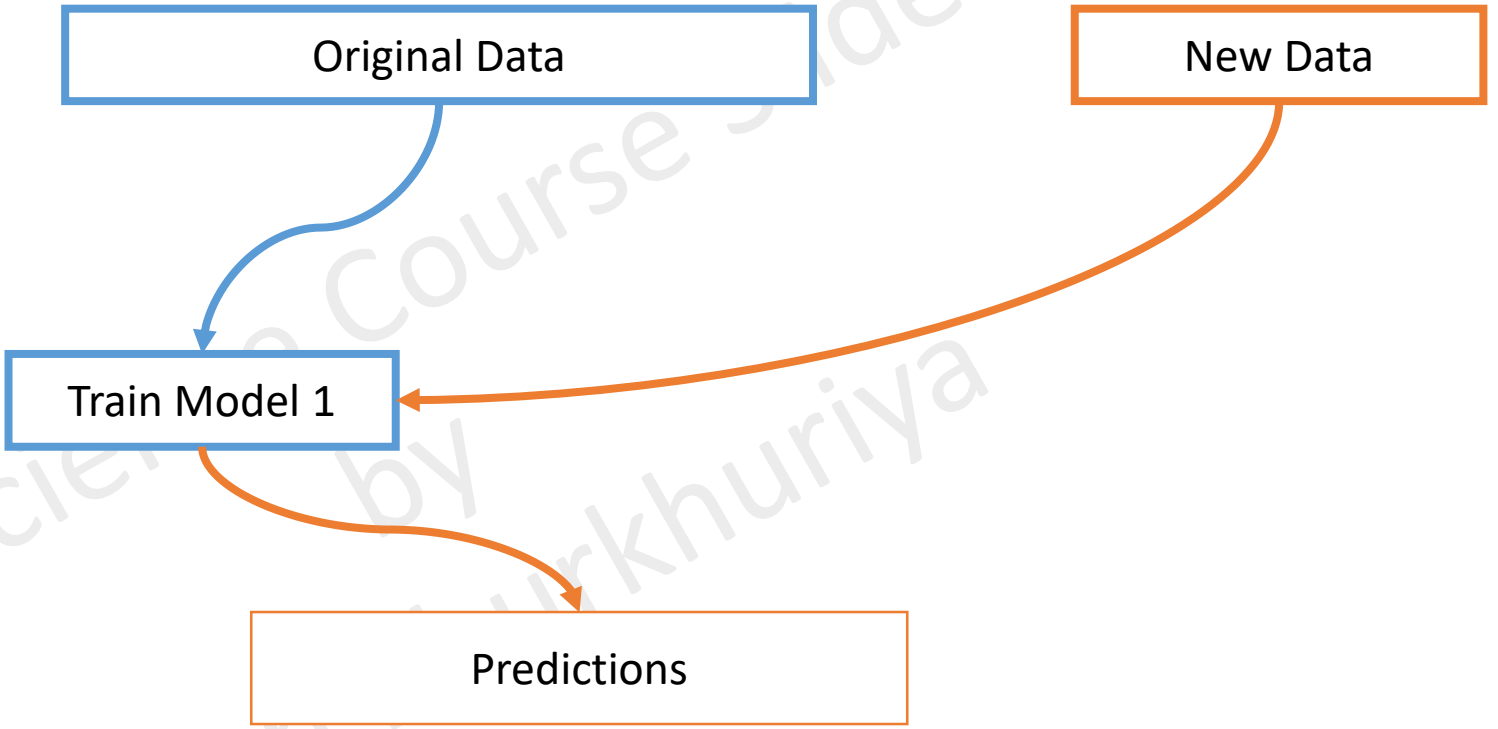
Train and Test Split



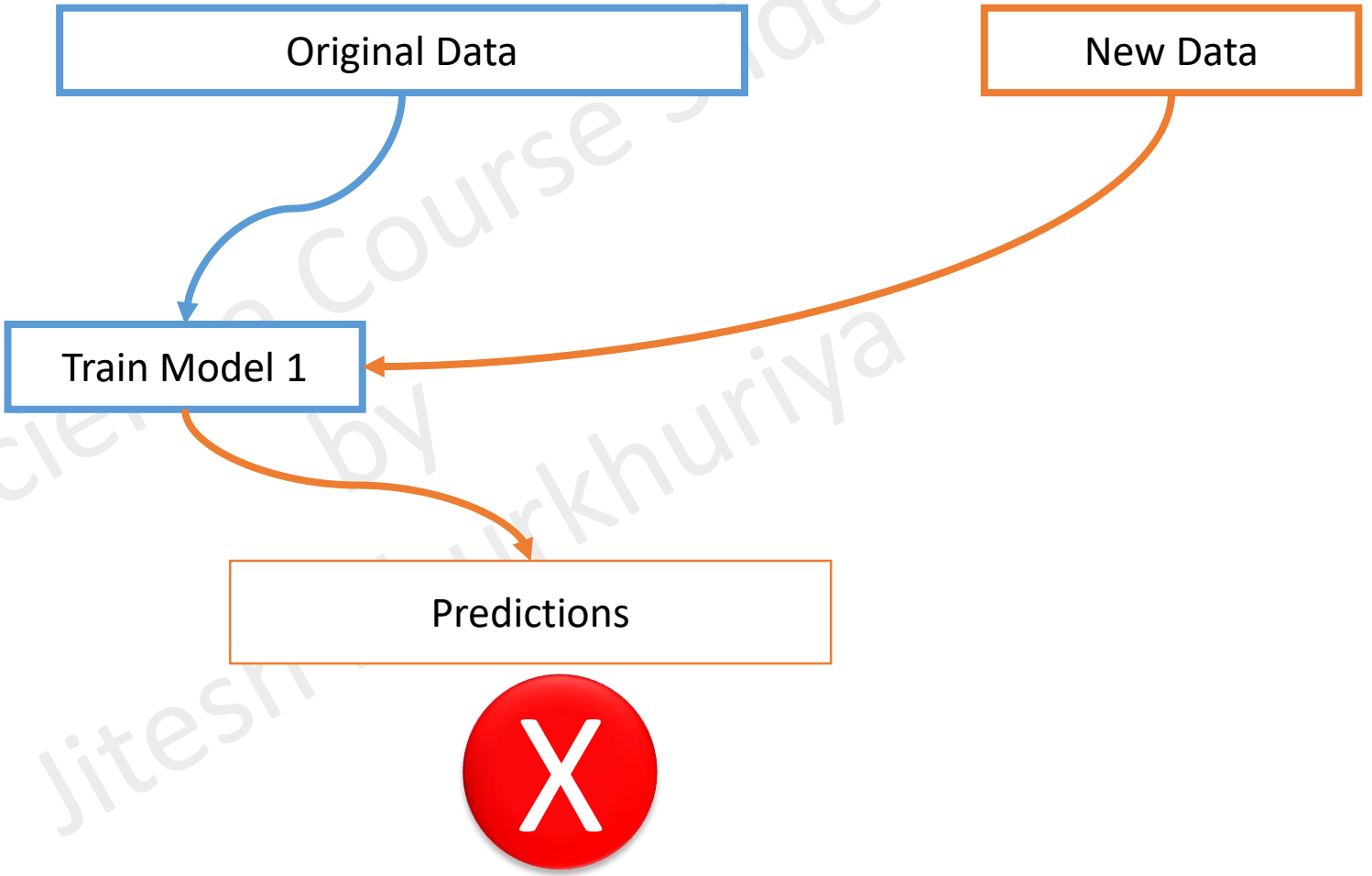
Why to Split the data?



Why to Split the data?

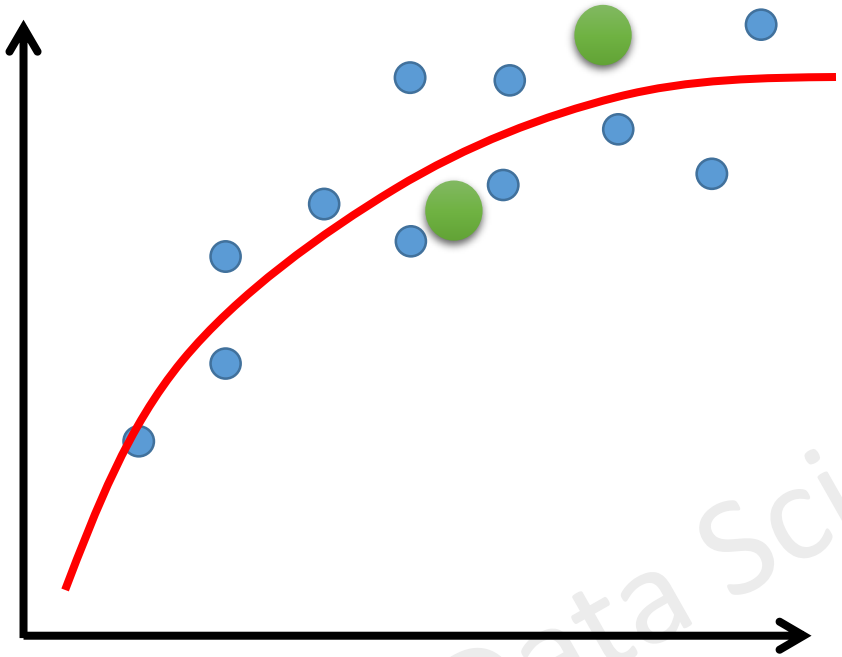


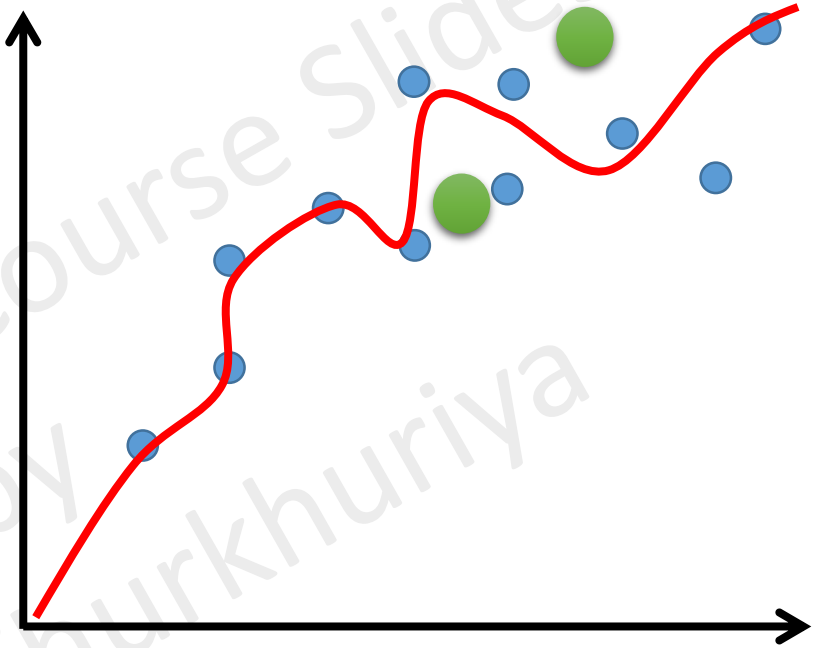
Why to Split the data?



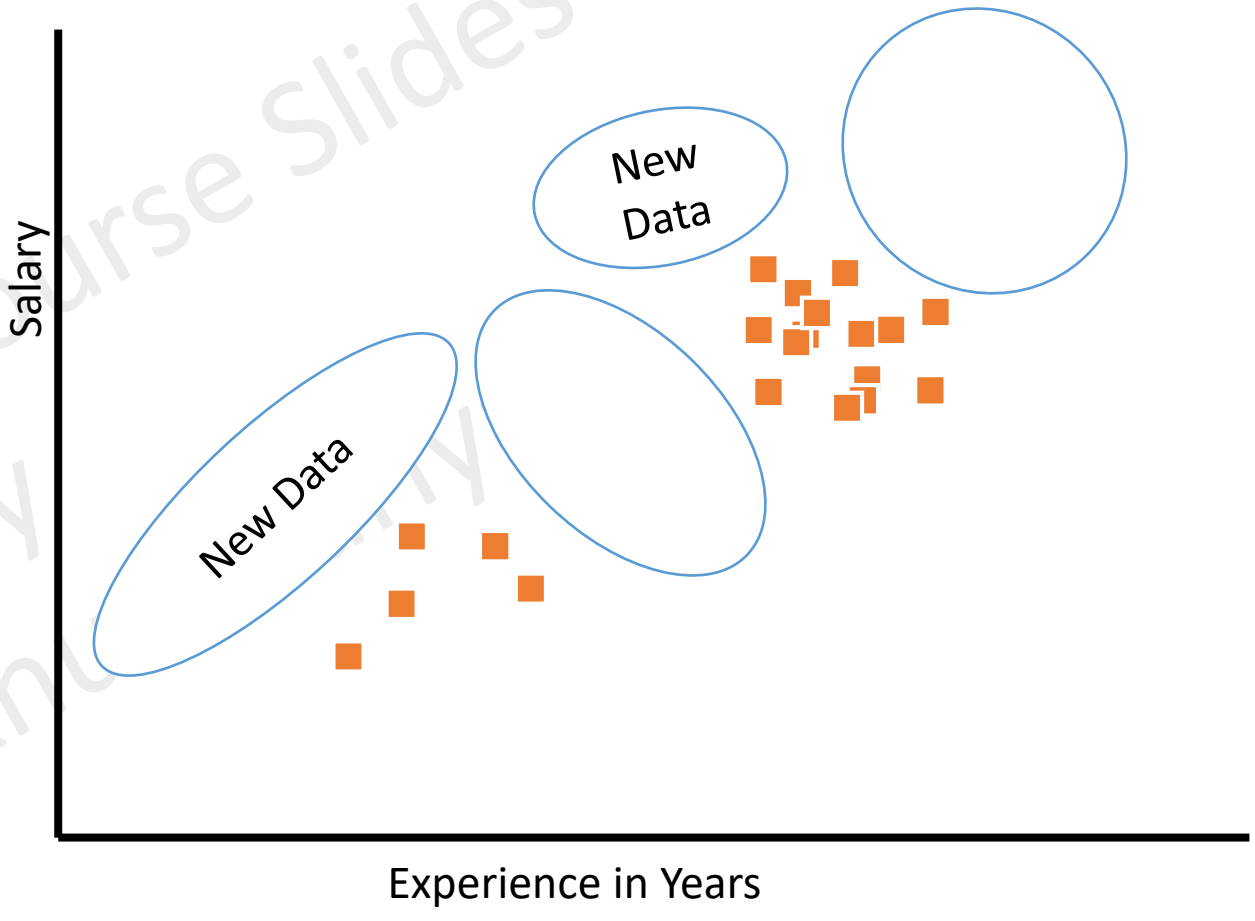
Why it may happen?

Problem of Overfitting

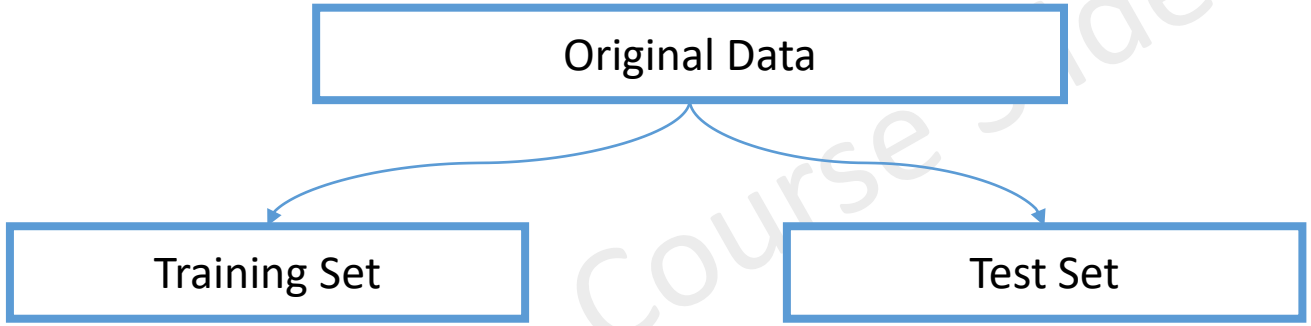




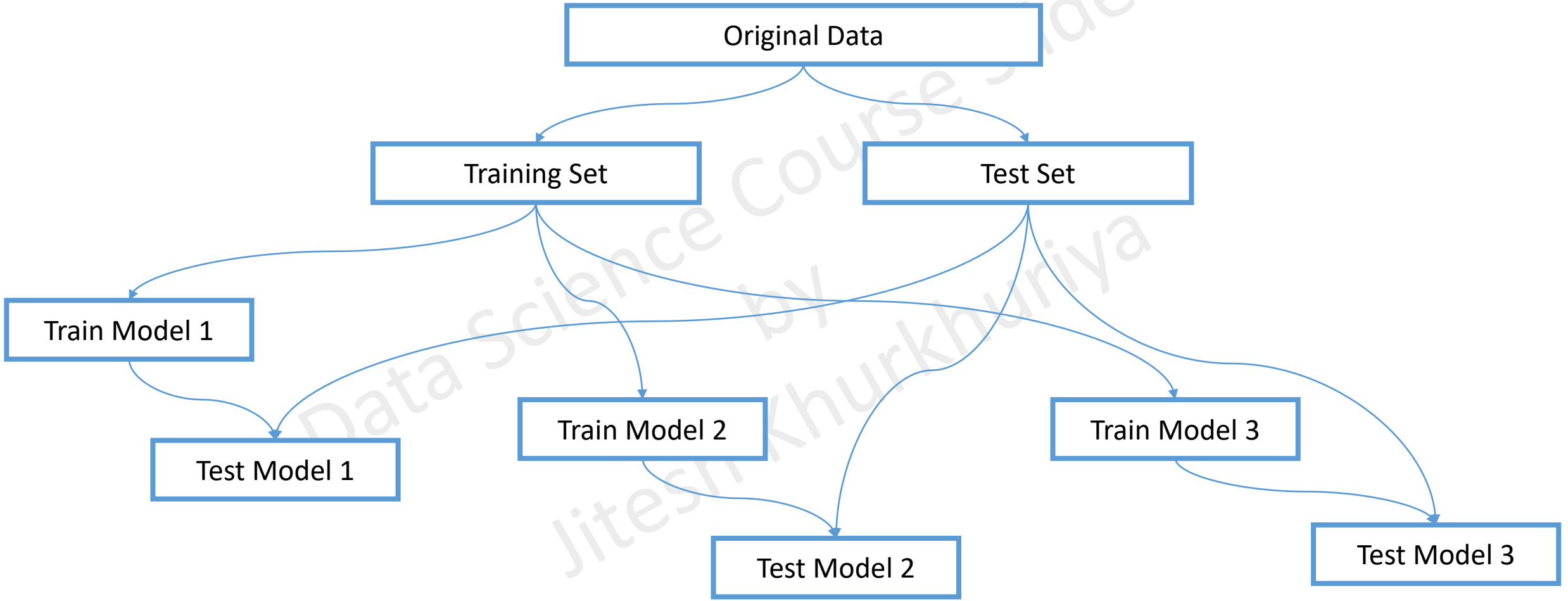
Sample or Selection Bias



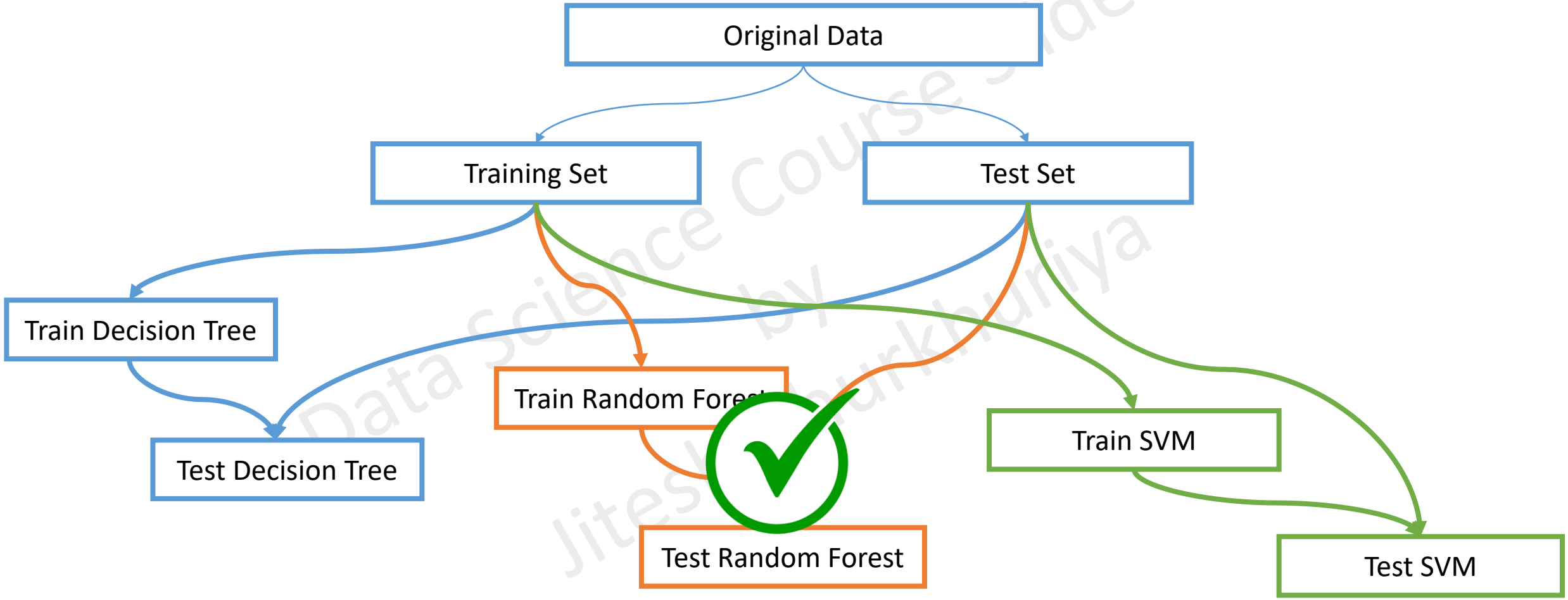
Why to Split the data?



Model Selection



Model Selection



Model Test Score for Adult Income Prediction

Split Random Seed	Split Size	Decision Tree	Random Forest	SVM
0	0.2	77.08%	79.18%	80.24%
123	0.2	78.39%	79.15%	**80.54%**
456	0.2	78.32%	78.57%	80.41%
999	0.2	**76.93%**	78.67%	79.73%
0	0.33	77.10%	79.30%	80.10%
123	0.33	77.81%	79.03%	79.46%
456	0.33	78.11%	79.31%	79.93%
999	0.33	77.70%	**78.39%**	79.49%
0	0.4	77.34%	78.96%	79.88%
123	0.4	78.44%	**79.87%**	79.63%
456	0.4	78.34%	79.01%	79.88%
999	0.4	77.43%	79.01%	79.79%
0	0.45	77.59%	79.30%	79.59%
123	0.45	78.06%	79.20%	**79.43%**
456	0.45	**78.50%**	79.29%	79.87%
999	0.45	77.20%	79.00%	79.71%

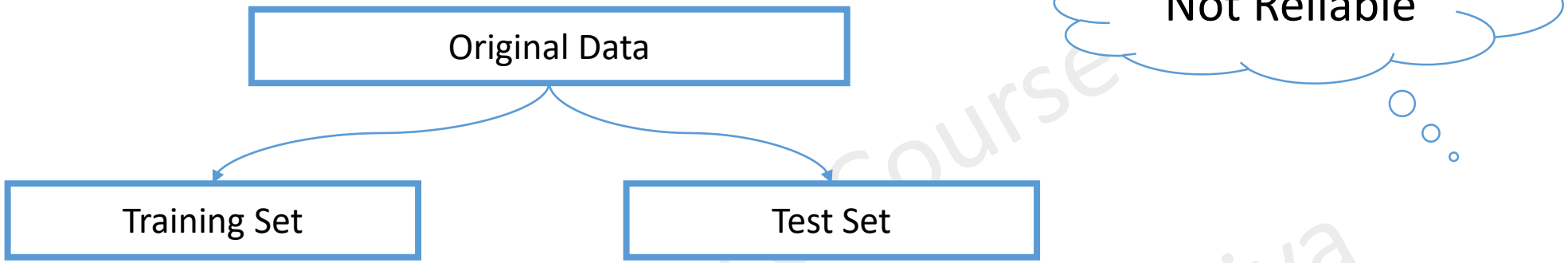
Model Test Score for Adult Income Prediction

Split Random Seed	Split Size	Decision Tree	Random Forest	SVM
0	0.2	77.08%	79.18%	80.24%
123	0.2	78.39%	79.15%	**80.54%**
456	0.2	78.32%	78.57%	80.41%
999	0.2	**76.93%**	78.67%	79.73%
0	0.33	77.10%	79.30%	80.10%
123	0.33	77.81%	79.03%	79.46%
456	0.33	78.11%	79.31%	79.93%
999	0.33	77.70%	**78.39%**	
0	0.4	77.34%	78.96%	
123	0.4	78.44%	**79.87%**	
456	0.4	78.34%	79.01%	79.88%
999	0.4	77.43%	79.01%	79.79%
0	0.45	77.59%	79.30%	79.59%
123	0.45	78.06%	79.20%	**79.43%**
456	0.45	**78.50%**	79.29%	79.87%
999	0.45	77.20%	79.00%	79.71%

Model Test Score for Adult Income Prediction

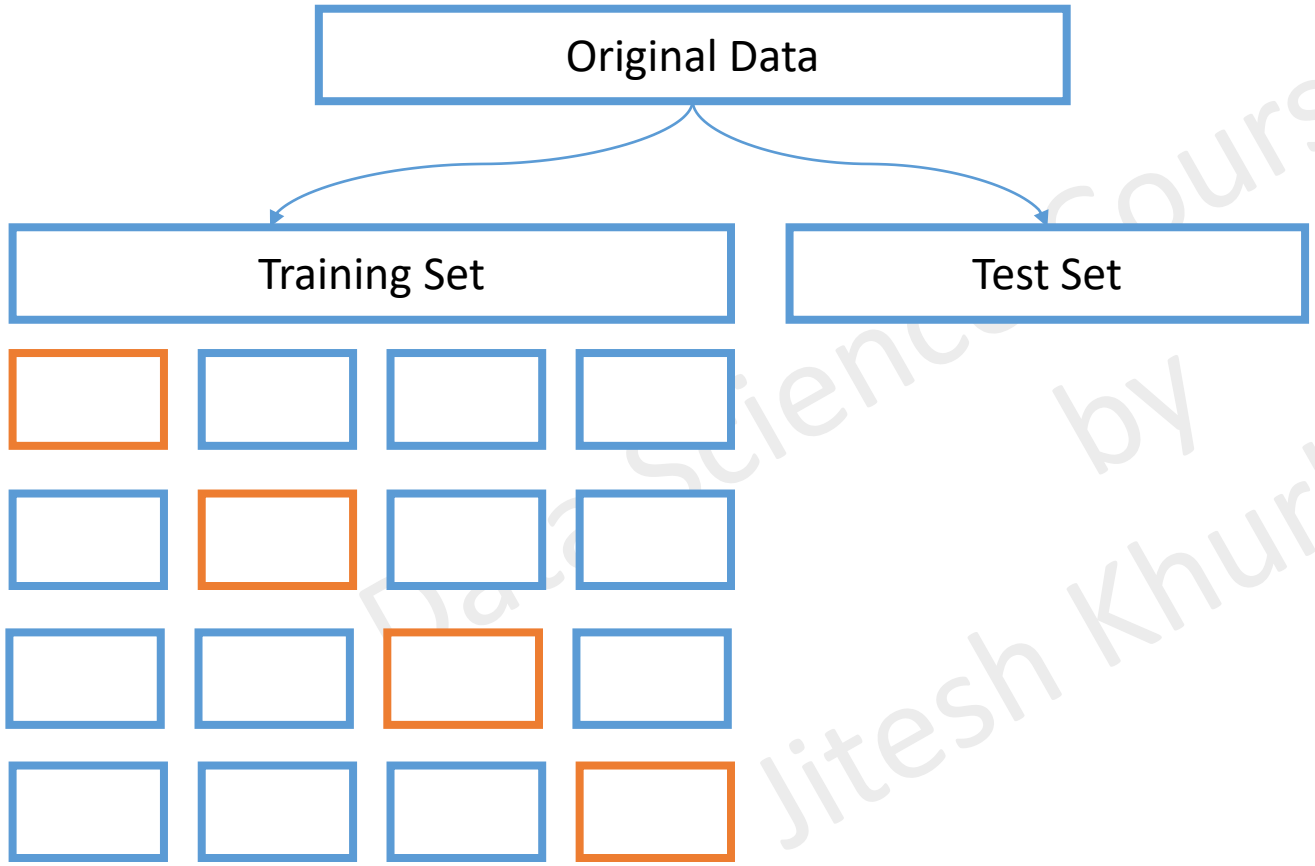
Split Random Seed	Split Size	Decision Tree	Random Forest	SVM
0	0.2	77.08%	79.18%	80.24%
123	0.2	78.39%	79.15%	**80.54%**
456	0.2	78.32%	78.57%	80.41%
999	0.2	**76.93%**	78.67%	79.73%
0	0.33	77.10%	79.30%	80.10%
123	0.33	77.81%	[illegible]	79.46%
456	0.33	78.11%	[illegible]	79.93%
999	0.33	77.70%	[illegible]	79.49%
0	0.4	77.34%	[illegible]	79.88%
123	**0.4**	**78.44%**	**79.87%**	**79.63%**
456	0.4	78.34%	79.01%	79.88%
999	0.4	77.43%	79.01%	79.79%
0	0.45	77.59%	79.30%	79.59%
123	0.45	78.06%	79.20%	**79.43%**
456	0.45	**78.50%**	79.29%	79.87%
999	0.45	77.20%	79.00%	79.71%

Model Selection

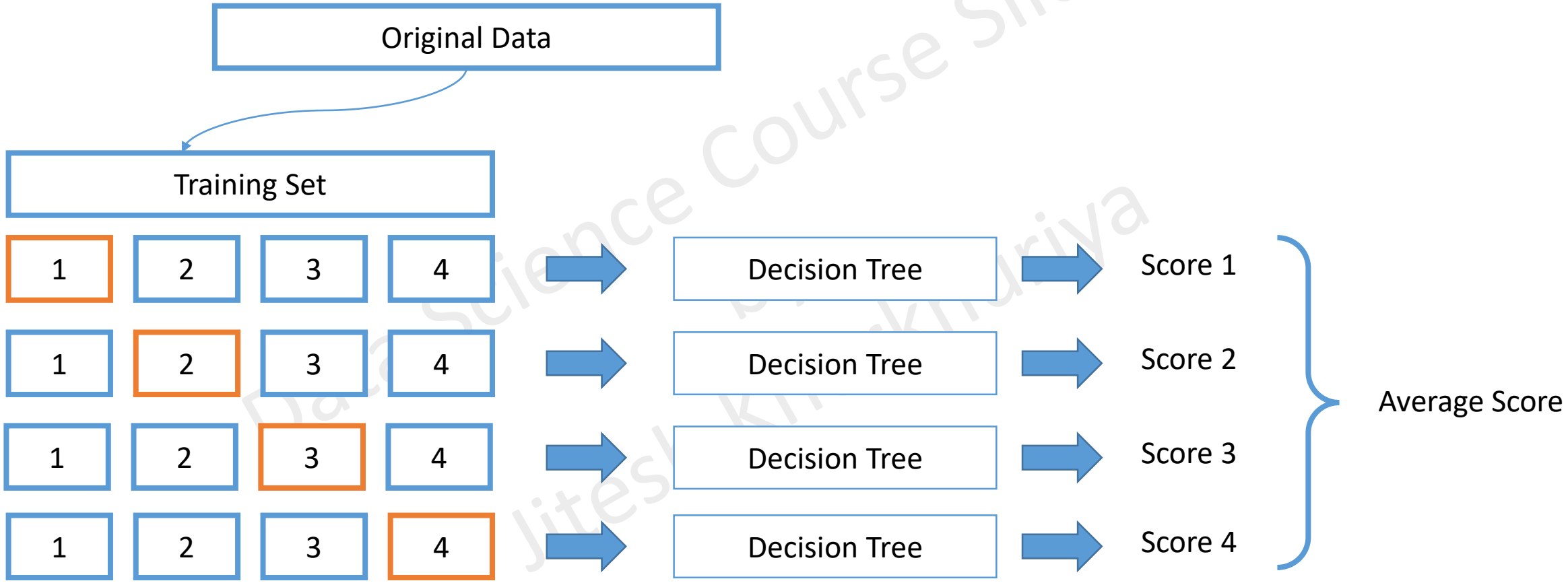


Not Reliable

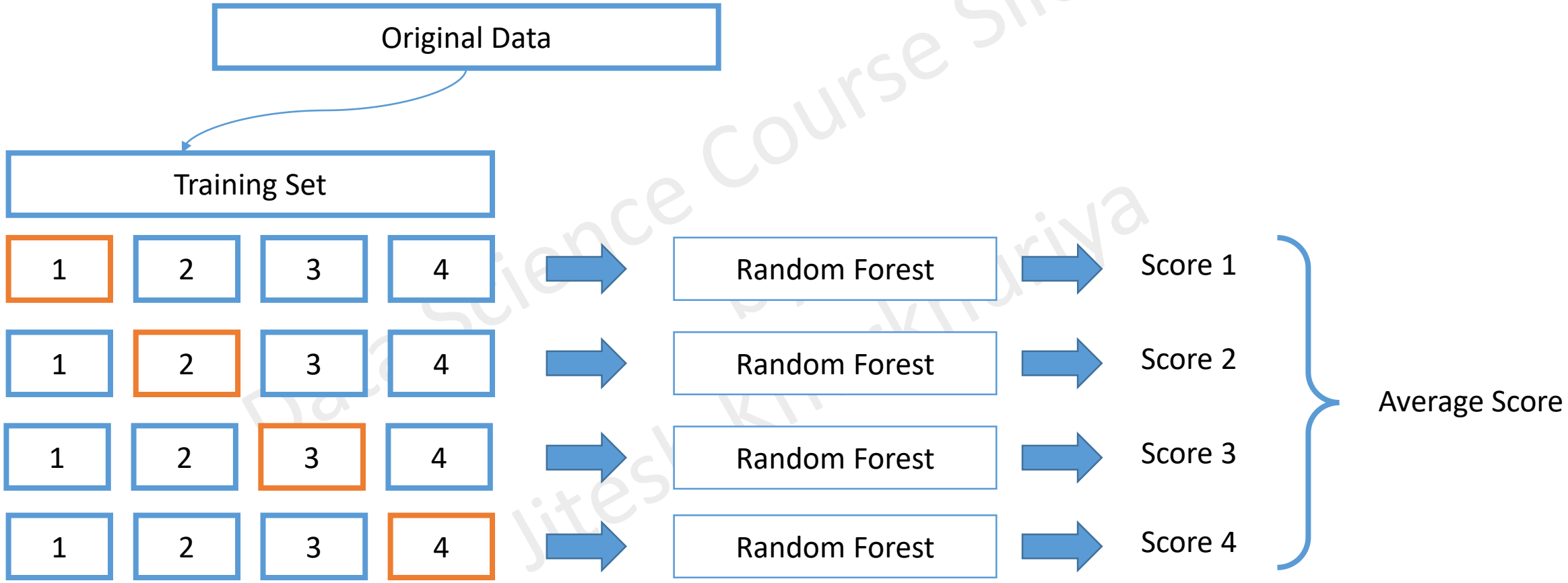
Cross Validation



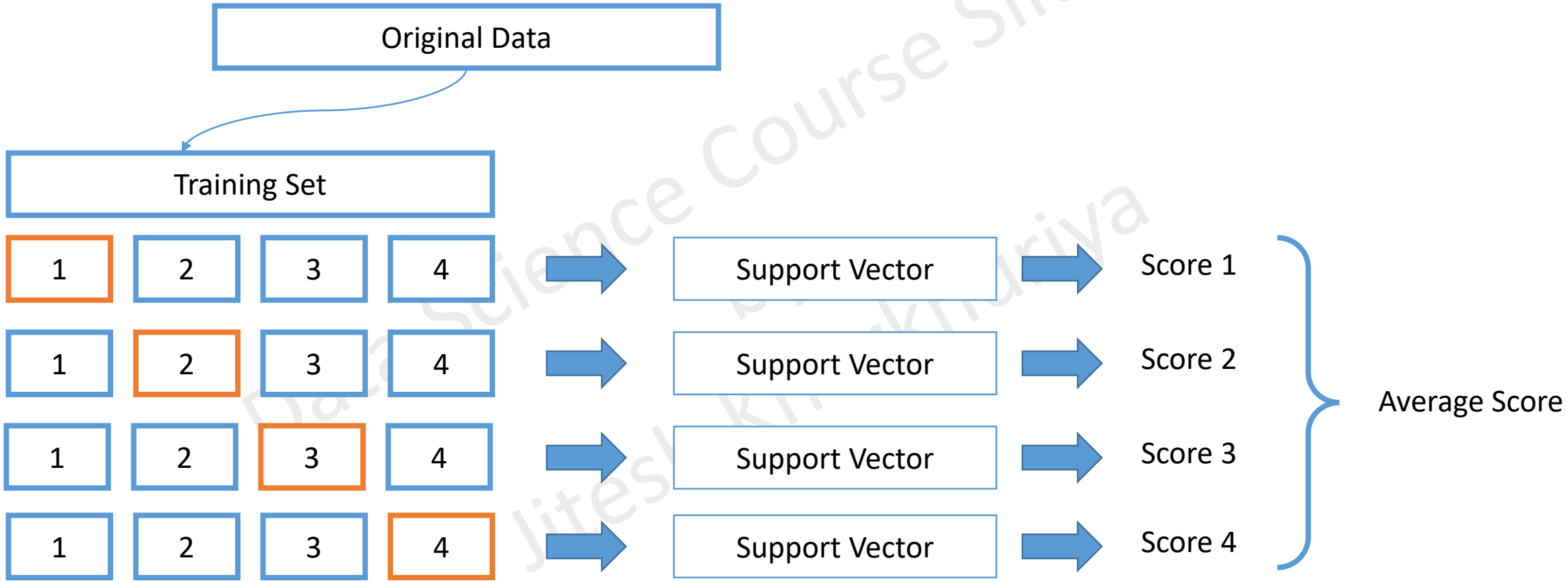
K-Fold Cross Validation



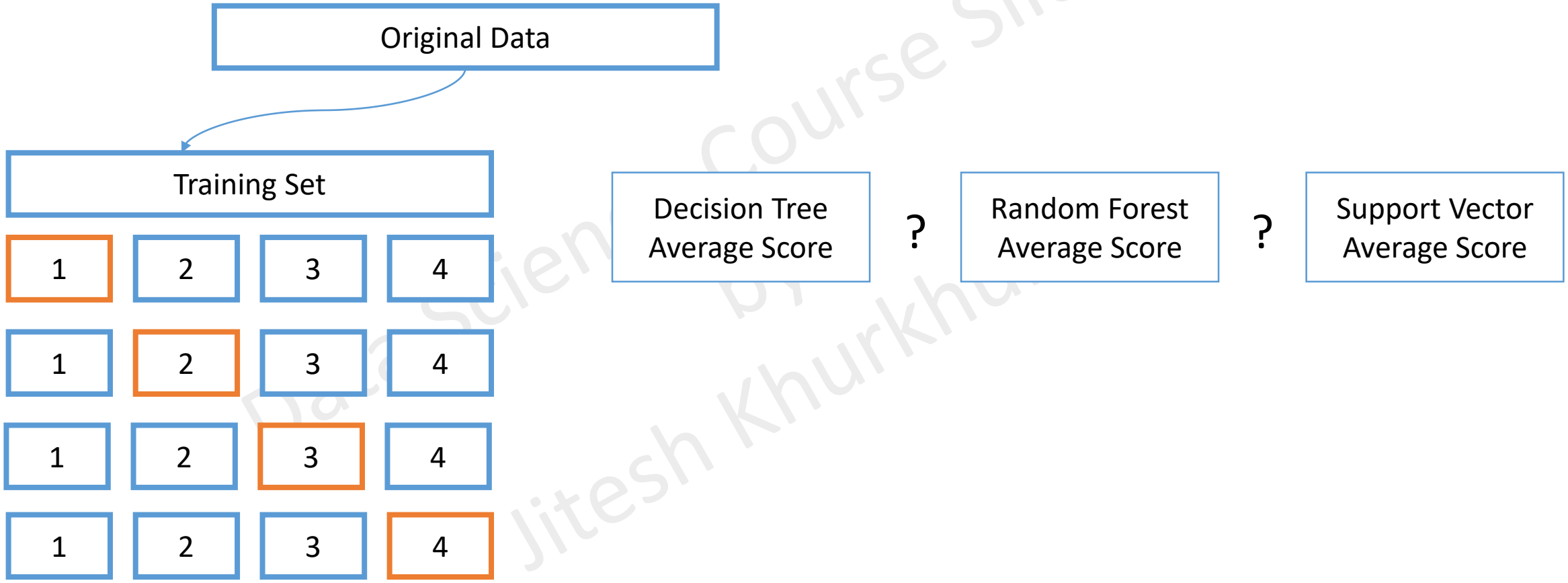
K-Fold Cross Validation



K-Fold Cross Validation



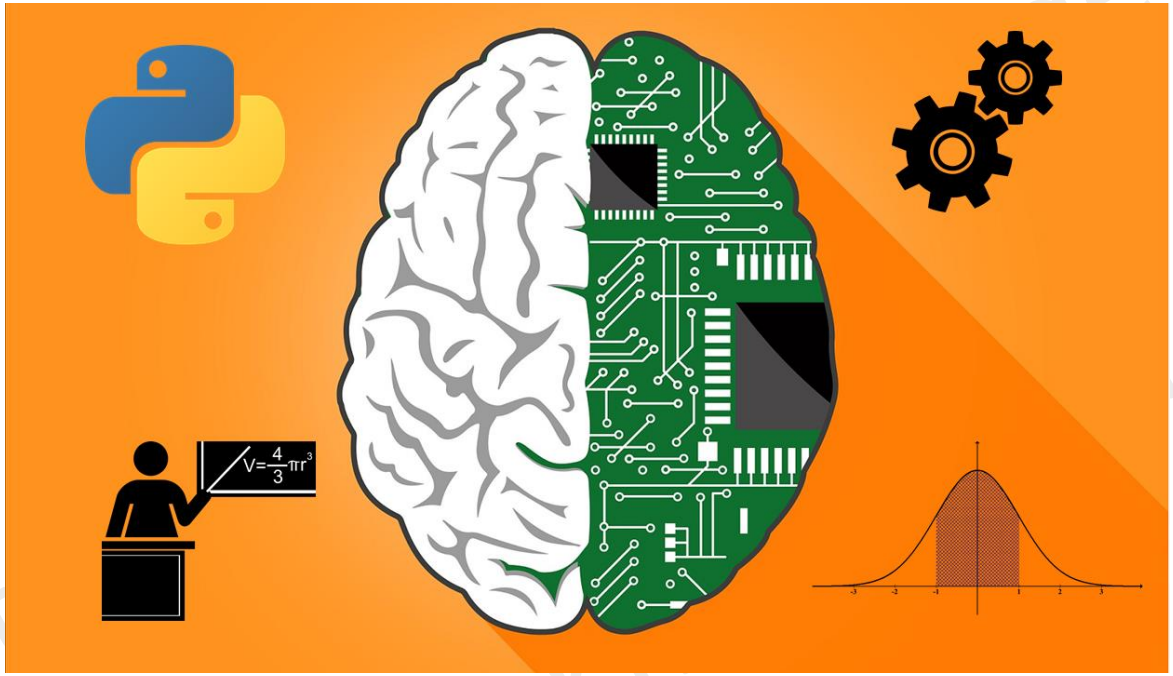
K-Fold Cross Validation



Random Sampling

Split Random Seed	Split Size	Decision Tree	Random Forest	SVM
0	0.2	77.08%	79.18%	80.24%
123	0.2	78.39%	79.15%	80.54%
456	0.2	78.32%	78.57%	80.41%
999	0.2	76.93%	78.67%	79.73%
	Average	**77.68%**	**78.89%**	**80.23%**
0	0.33	77.10%	79.30%	80.10%
123	0.33	77.81%	79.03%	79.46%
456	0.33	78.11%	79.31%	79.93%
999	0.33	77.70%	78.39%	79.49%
	Average	**77.68%**	**79.01%**	**79.75%**
0	0.4	77.34%	78.96%	79.88%
123	0.4	78.44%	**79.87%**	79.63%
456	0.4	78.34%	79.01%	79.88%
999	0.4	77.43%	79.01%	79.79%
	Average	**77.89%**	**79.21%**	**79.80%**

Complete Data Science and Machine Learning Using Python



Thank You!