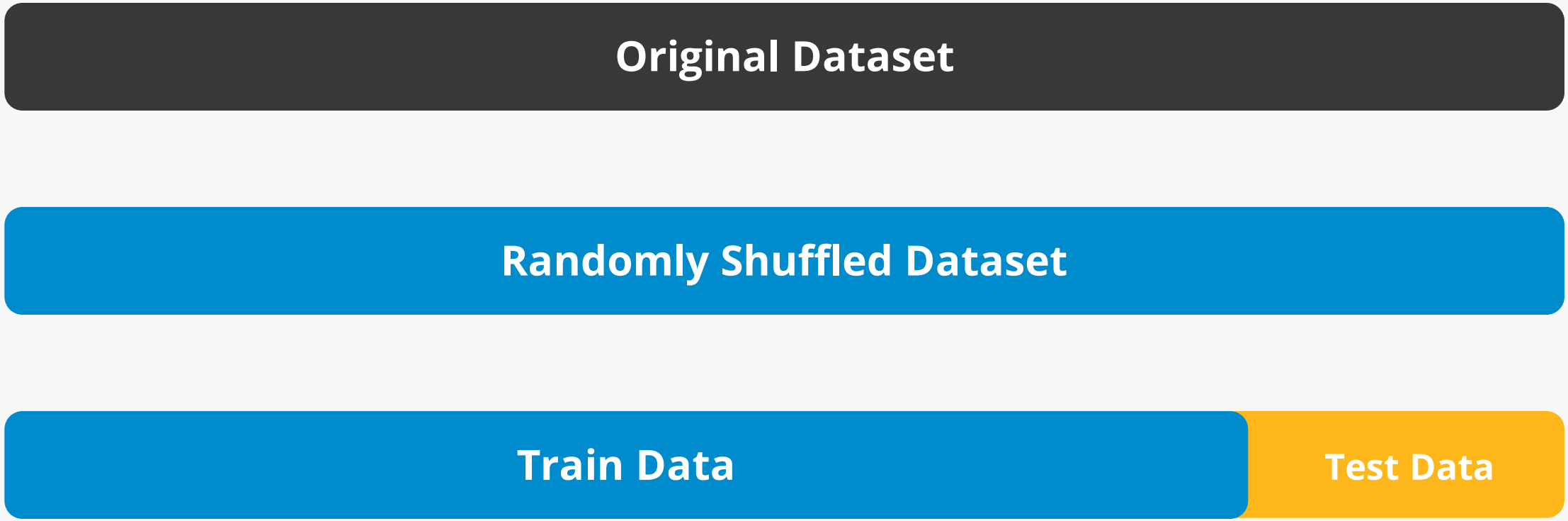
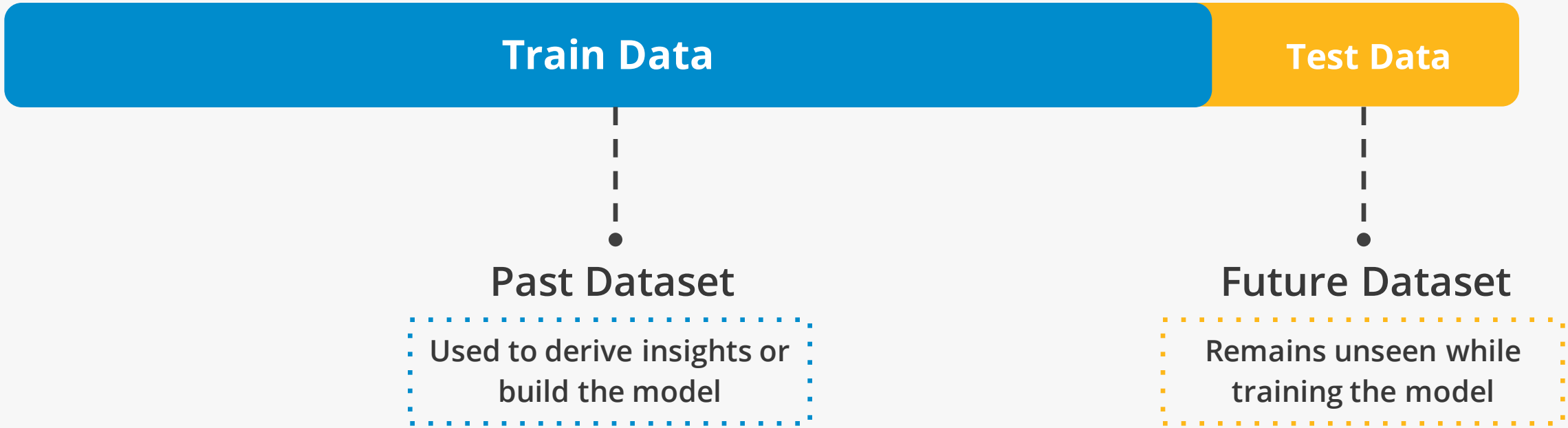
Train-Test Split

Train-Test Split

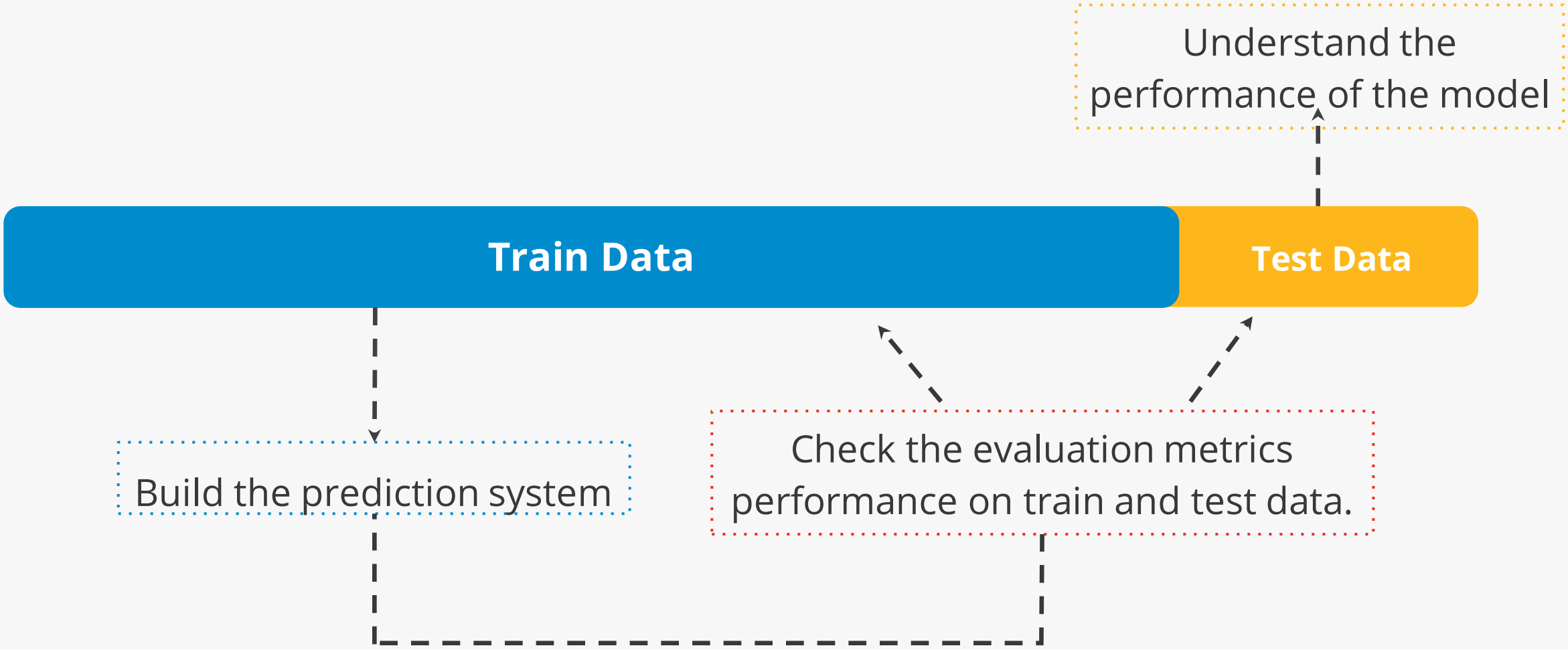


Why divide the dataset into train and test before training the model?

Train-Test Split

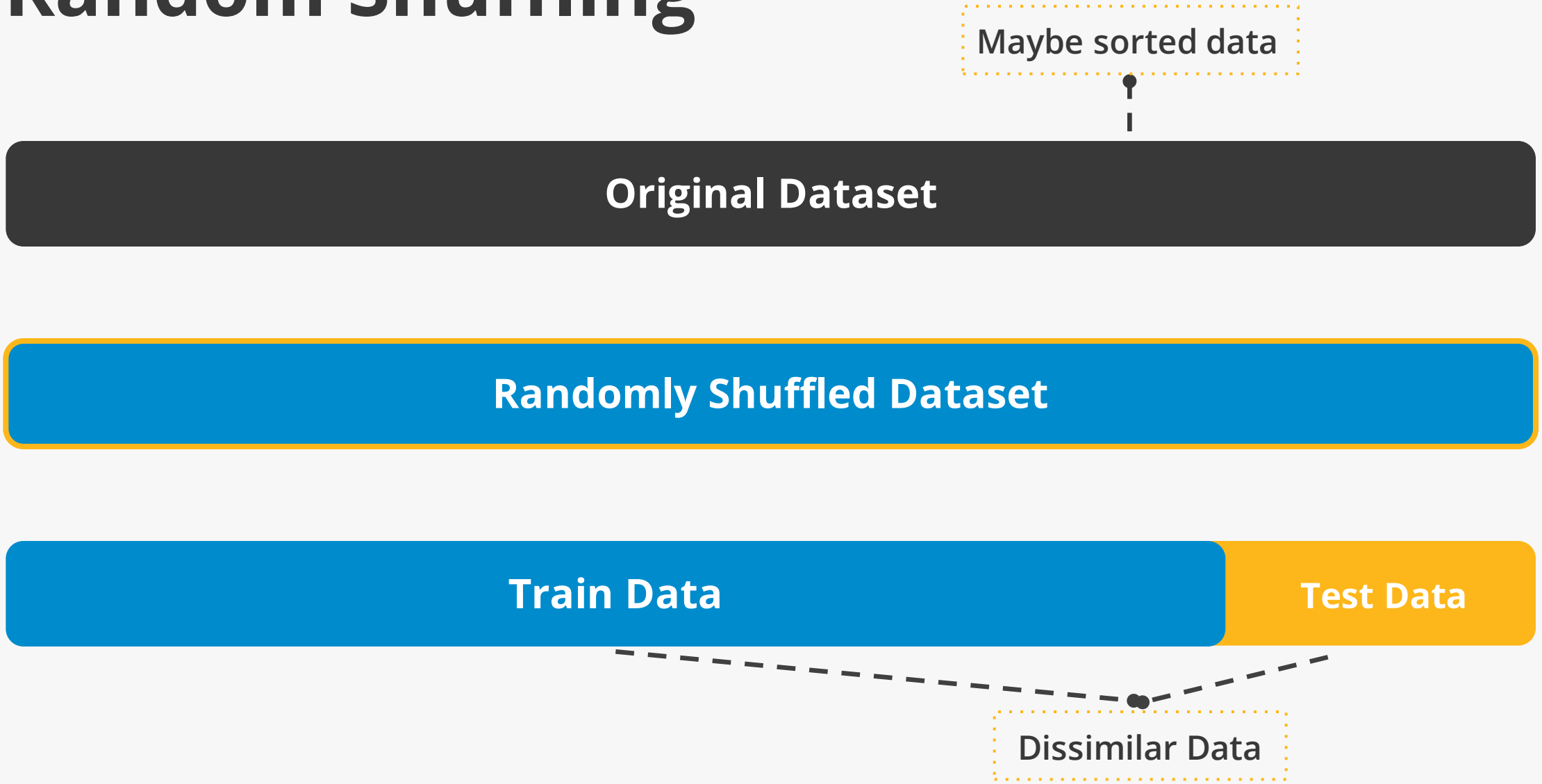


Train-Test Split

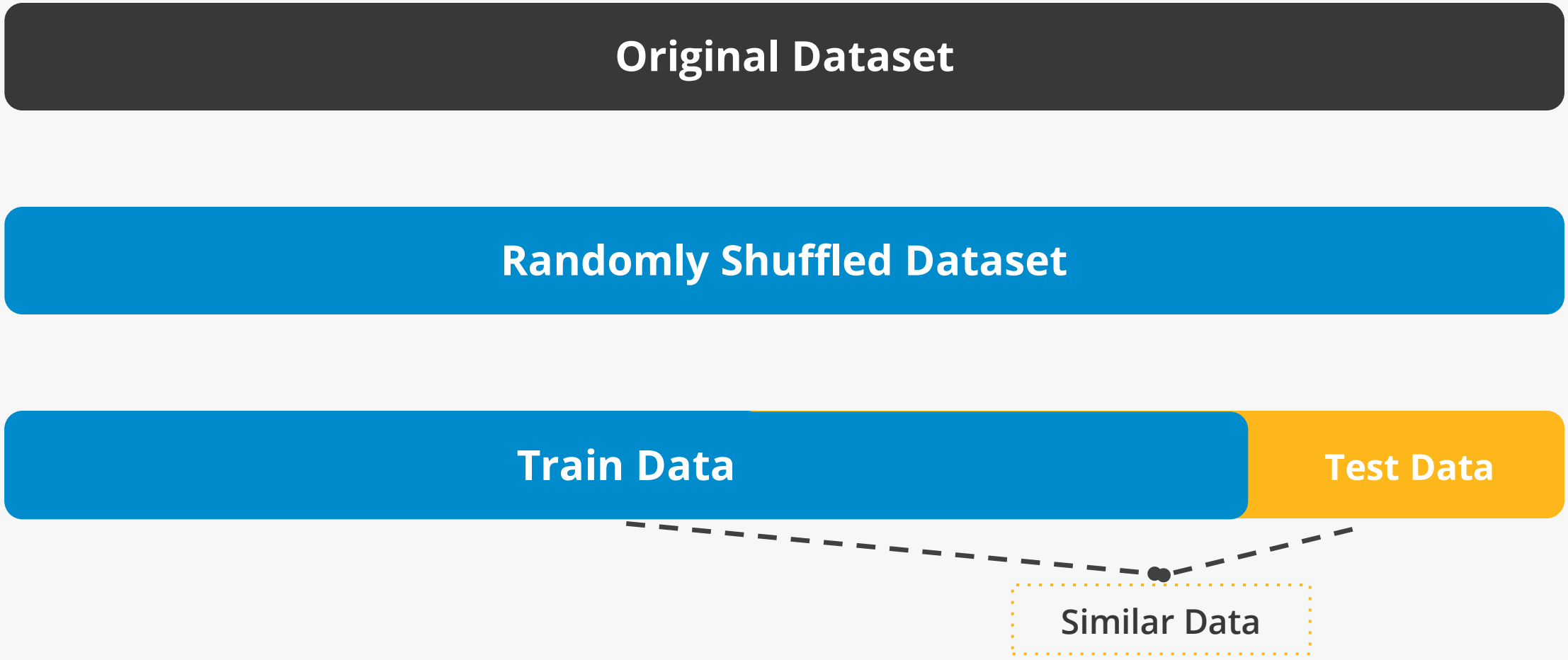


Random Shuffling

Random Shuffling



Random Shuffling



Random Shuffling: Example

Random Shuffling: Example

Client ID	Type of Client	Number of Elevators Required	Type of Building	Floor Area of Elevators (sq ft)	Number of Floors	Buy (Target)
1	Small	5	Mall	30	20	No
2	Medium	10	Mall	50	30	Yes
3	Medium	8	Mall	40	20	No
.....	.....	.....	.....	.....	.....	.....
500	Big	50	Office	35	5	Yes
501	Medium	20	Office	75	6	No
.....	.....	.....	.....	.....	.....	.....
998	Medium	20	Residential	75	6	No
999	Big	25	Residential	50	10	No
1000	Small	3	Residential	25	10	Yes

Random Shuffling: Example

Client ID	Type of Client	Number of Elevators Required	Type of Building	Floor Area of Elevators (sq ft)	Number of Floors	Buy (Target)
1	Small	5	Mall	30	20	No
2	Medium	10	Mall	50	30	Yes
3	Medium	8	Mall	40	20	No
.....	.....	.....	.....	.....	.....	.....
500	Big	50	Office	35	5	Yes
501	Medium	20	Office	75	6	No
.....	.....	.....	.....	.....	.....	.....
998	Medium	20	Residential	75	6	No
999	Big	25	Residential	50	10	No
1000	Small	3	Residential	25	10	Yes

Random Shuffling: Example

Client ID	Type of Client	Number of Elevators Required	Type of Building	Floor Area of Elevators (sq ft)	Number of Floors	Buy (Target)
1	Small	5	Mall	30	20	No
2	Medium	10	Mall	50	30	Yes
3	Medium	8	Mall	40	20	No
.....	.....	.....	.....	.....	.....	.....
500	Big	50	Office	35	5	Yes
501	Medium	20	Office	75	6	No
.....	.....	.....	.....	.....	.....	.....
998	Medium	20	Residential	75	6	No
999	Big	25	Residential	50	10	No
1000	Small	3	Residential	25	10	Yes

Random Shuffling: Example

Client ID	Type of Client	Number of Elevators Required	Type of Building	Floor Area of Elevators (sq ft)	Number of Floors	Buy (Target)
1	Small	5	Mall	30	20	No
2	Medium	10	Mall	50	30	Yes
3	Medium	8	Mall	40	20	No
.....	.....	.....	.....	.....	.....	.....
500	Big	50	Office	35	5	Yes
501	Medium	20	Office	75	6	No
.....	.....	.....	.....	.....	.....	.....
998	Medium	20	Residential	75	6	No
999	Big	25	Residential	50	10	No
1000	Small	3	Residential	25	10	Yes

Random Shuffling: Example

	Client ID	Type of Client	Number of Elevators Required	Type of Building	Floor Area of Elevators (sq ft)	Number of Floors	Buy (Target)
T	1	Small	5	Mall	30	20	No
R	2	Medium	10	Mall	50	30	Yes
	3	Medium	8	Mall	40	20	No
A	.....	.....	.....	.....	.....	.....	.....
	500	Big	50	Office	35	5	Yes
I	501	Medium	20	Office	75	6	No
N	.....	.....	.....	.....	.....	.....	.....
	998	Medium	20	Residential	75	6	No
	999	Big	25	Residential	50	10	No
	1000	Small	3	Residential	25	10	Yes

Random Shuffling: Example

	Client ID	Type of Client	Number of Elevators Required	Type of Building	Floor Area of Elevators (sq ft)	Number of Floors	Buy (Target)
	1	Small	5	Mall	30	20	No
	2	Medium	10	Mall	50	30	Yes
	3	Medium	8	Mall	40	20	No
	.....	.....	.....	.....	.....	.....	.....
	500	Big	50	Office	35	5	Yes
	501	Medium	20	Office	75	6	No
TEST	.....	.....	.....	.....	.....	.....	.....
	998	Medium	20	Residential	75	6	No
	999	Big	25	Residential	50	10	No
	1000	Small	3	Residential	25	10	Yes

Random Shuffling: Example

	Client ID	Type of Client	Number of Elevators Required	Type of Building	Floor Area of Elevators (sq ft)	Number of Floors	Buy (Target)
T	1	Small	5	Mall	30	20	No
R	2	Medium	10	Mall	50	30	Yes
	3	Medium	8	Mall	40	20	No
A	.....	.....	.....	.....	.....	.....	.....
	500	Big	50	Office	35	5	Yes
I	501	Medium	20	Office	75	6	No
N	.....	.....	.....	.....	.....	.....	.....
	998	Medium	20	Residential	75	6	No
	999	Big	25	Residential	50	10	No
	1000	Small	3	Residential	25	10	Yes

Random Shuffling: Example

	Client ID	Type of Client	Number of Elevators Required	Type of Building	Floor Area of Elevators (sq ft)	Number of Floors	Buy (Target)
	1	Small	5	Mall	30	20	No
	2	Medium	10	Mall	50	30	Yes
	3	Medium	8	Mall	40	20	No
	.....	.....	.....	.....	.....	.....	.....
	500	Big	50	Office	35	5	Yes
	501	Medium	20	Office	75	6	No
TEST	.....	.....	.....	.....	.....	.....	.....
	998	Medium	20	**Residential**	75	6	No
	999	Big	25	**Residential**	50	10	No
	1000	Small	3	**Residential**	25	10	Yes

Random Shuffling: Example

Client ID	Type of Client	Number of Elevators Required	Type of Building	Floor Area of Elevators (sq ft)	Number of Floors	Buy (Target)
1000	Small	3	Residential	25	10	Yes
2	Medium	10	Mall	50	30	Yes
999	Big	25	Residential	50	10	No
.....	.....	.....	.....	.....	.....	.....
400	Big	50	Office	35	5	Yes
3	Medium	8	Mall	40	20	No
.....	.....	.....	.....	.....	.....	.....
1	Small	5	Mall	30	20	No
401	Medium	20	Office	75	6	No
998	Medium	20	Residential	75	6	No

Random Shuffling: Example

	Client ID	Type of Client	Number of Elevators Required	Type of Building	Floor Area of Elevators (sq ft)	Number of Floors	Buy (Target)
T	1000	Small	3	Residential	25	10	Yes
R	2	Medium	10	Mall	50	30	Yes
	999	Big	25	Residential	50	10	No
A	.....	.....	.....	.....	.....	.....	.....
	400	Big	50	Office	35	5	Yes
I	3	Medium	8	Mall	40	20	No
N	.....	.....	.....	.....	.....	.....	.....
	1	Small	5	Mall	30	20	No
	401	Medium	20	Office	75	6	No
	998	Medium	20	Residential	75	6	No

Random Shuffling: Example

Client ID	Type of Client	Number of Elevators Required	Type of Building	Floor Area of Elevators (sq ft)	Number of Floors	Buy (Target)
1000	Small	3	Residential	25	10	Yes
2	Medium	10	Mall	50	30	Yes
999	Big	25	Residential	50	10	No
.....	.....	.....	.....	.....	.....	.....
400	Big	50	Office	35	5	Yes
3	Medium	8	Mall	40	20	No
.....	.....	.....	.....	.....	.....	.....
1	Small	5	Mall	30	20	No
401	Medium	20	Office	75	6	No
998	Medium	20	Residential	75	6	No

TEST

Random Shuffling is not a suitable practice for time-series problems.

Explanation

Predict Future using Past Information

Shuffled Time Series Data

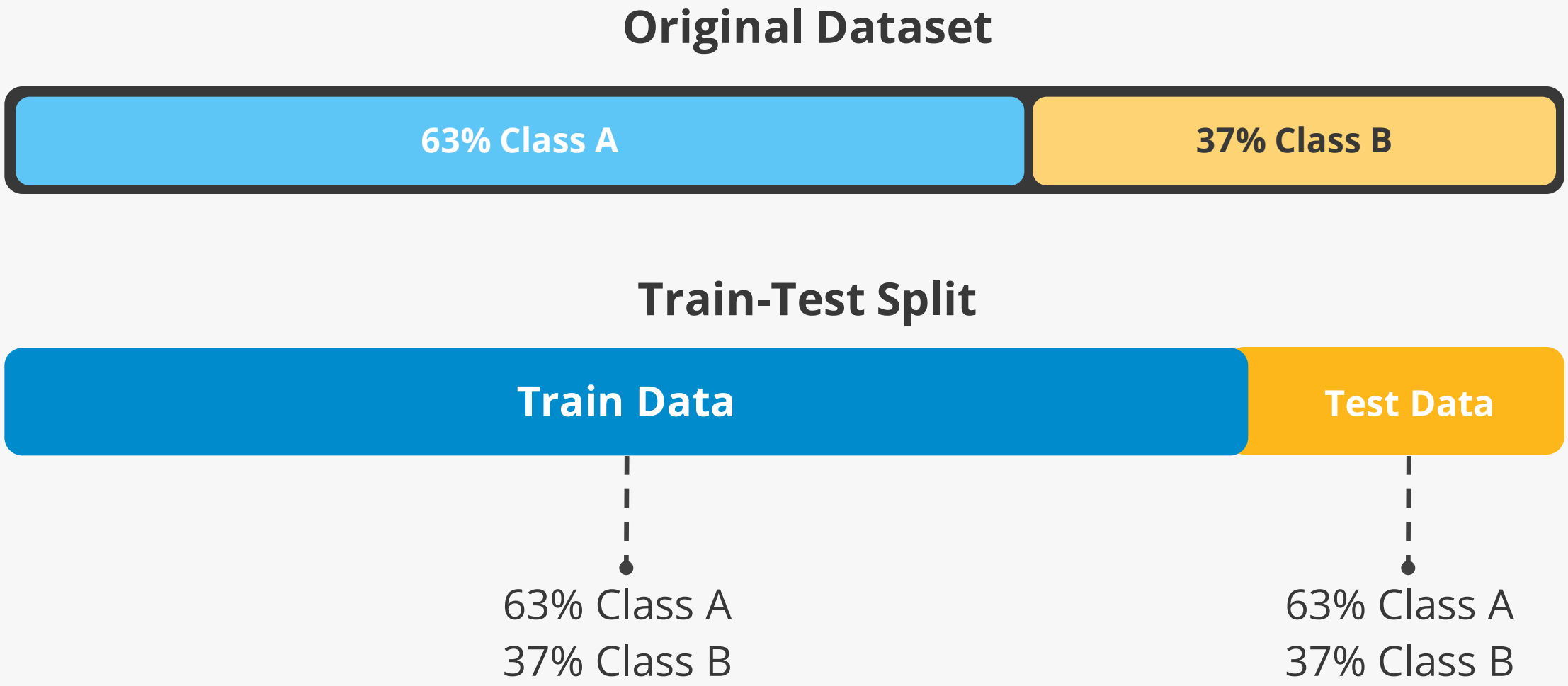
Good results on Train and Test

but

Bad results in real time scenarios

Stratified Sampling

Stratified Sampling



Implementing the concepts

Implementing Train-Test Split

```
#import train test split
from sklearn.model_selection import train_test_split

#split the data into train and test
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.3, random_state = 42)
```

Train-Test Ratio