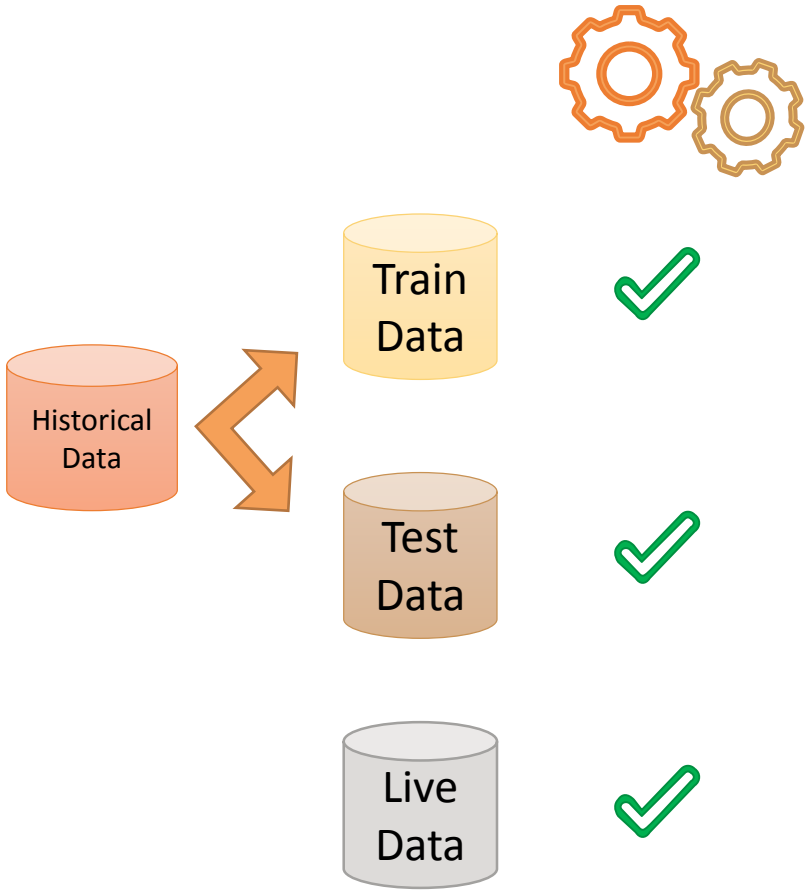
Train In Data

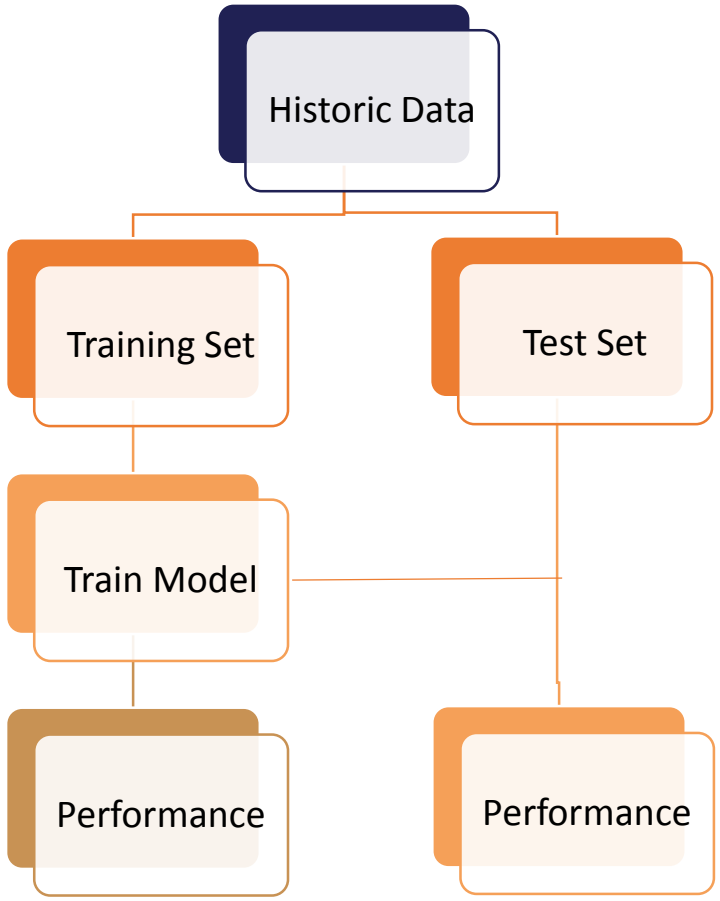
# How to set up a classifier with over-sampling

# Generalization




- Generalization is the ability of an algorithm to be effective across datasets
- We are particularly interested in the performance in real / live data
- Live data has a class imbalance
- Important to determine performance in a data with the class imbalance ➔ Test set

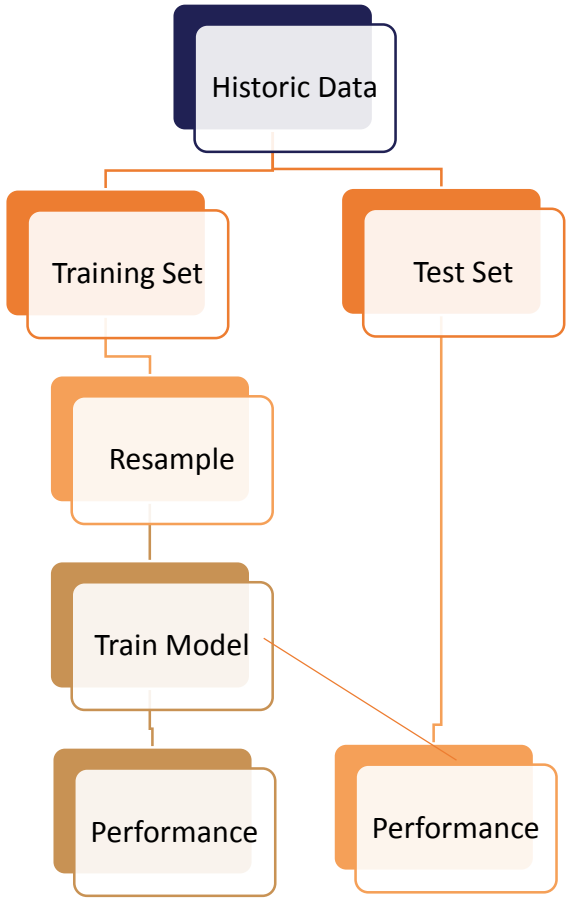
# Training a Machine Learning Model




To prevent over-fitting, it is common practice to:

- Separate the data into a train and a test set.
- Train the model in the train set
- Evaluate in the test set

# Training a Model with re-sampling

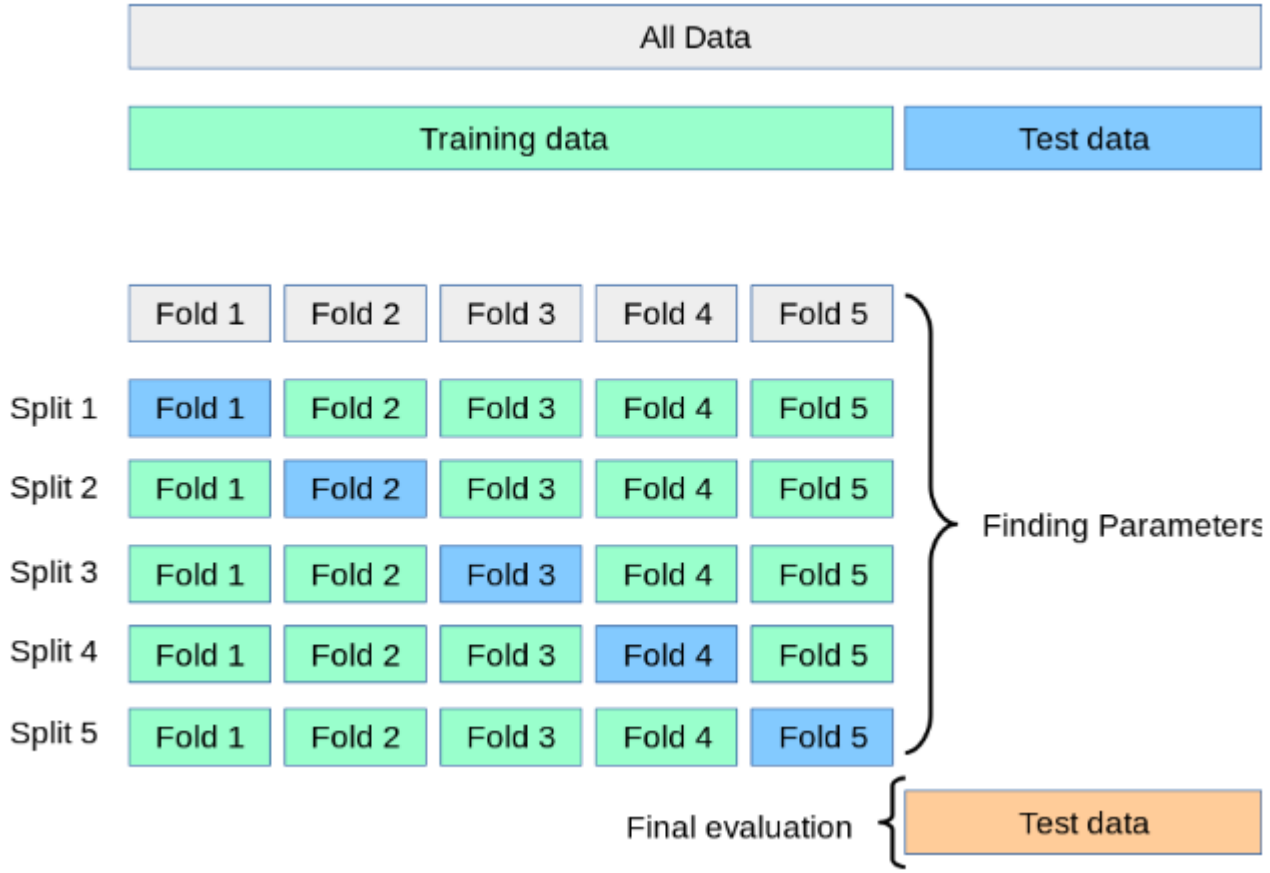



To prevent over-fitting, it is common practice to:

- Separate the data into a train and a test set.
- Re-sample the train set **only**
- Train the model in the resampled train set
- Evaluate in the test set

# Cross-Validation

- Train set divided into k folds
- Re-sample k-1 fold
- Train Model in resampled k-1 fold
- Tes Model in the $k^{th}$ fold (not resampled)




https://scikit-learn.org/stable/modules/cross_validation.html

Train In Data

# THANK YOU

www.trainindata.com