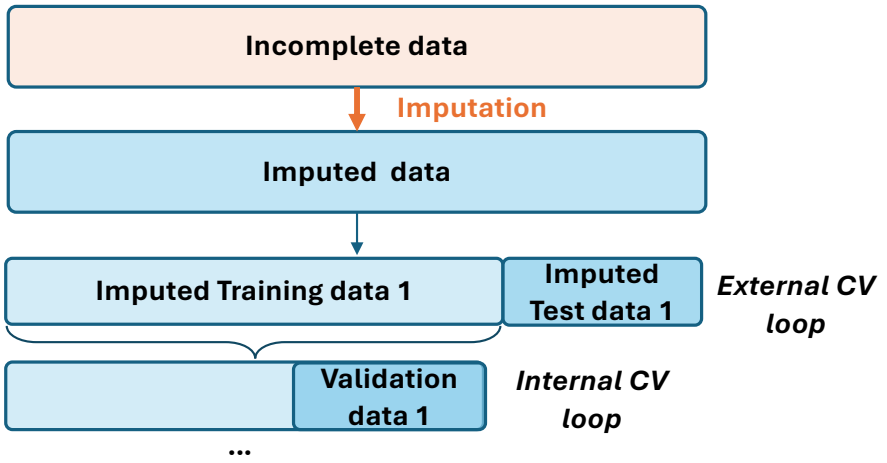
Option 1 (impute = 1): Split, then impute

Training data is imputed separately.

Option 1a: Test data is imputed by imputer model from training set (mean, K-NN)

Option 1b: Test data is imputed by using all available data (missForest)



Option 2 (impute = 2) : Impute, then split

NB: there is some spillage from test into train from imputation, but it is faster than Option 1.

! Warning: This option is intended for pilot investigations, or when the dataset is large, missingness is low, and computational power is not enough for Option 1.

Experiments show that the **performance estimates are biased up.**