Data split best practice:

Labeled data	Train	Val	Test

What %? Depends on size of data and amount of noise.

Smaller → higher variance

larger → less data for other splits

For small data, take full advantage of as much data as possible:

Val_1	Val_2	Val_3	...	Val_k	Test

"k-fold cross-validation":

train on each subset of k-1 chunks, val on the last

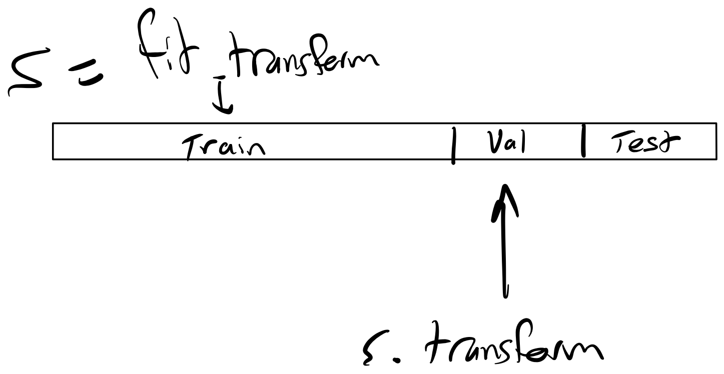
avg val accuracy across all k trials

+ better training, lower-variance val accuracy

- need to train k times

"leave-one-out cross-validation":

$k = n$



S = fit_transform([train])

S. transform(val)

S. transform(test)