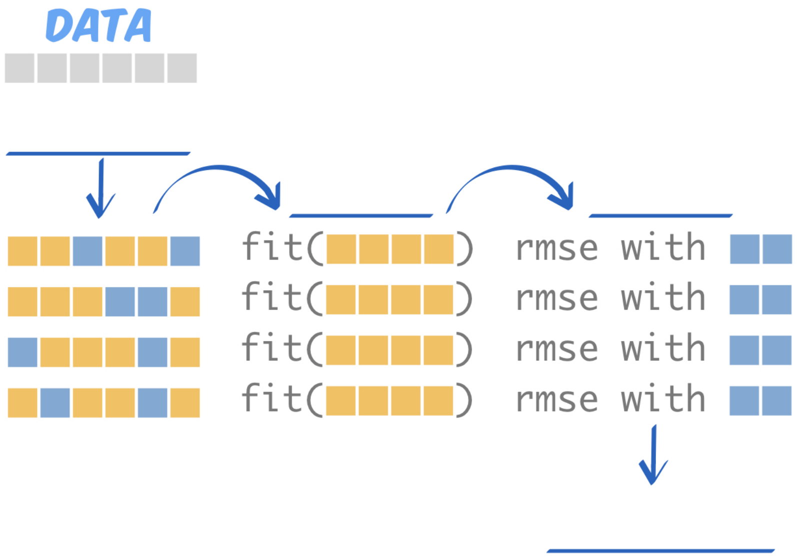
Cross-validation

Handout 5 of Introduction to Machine Learning

January 2019

1. Fill in the blanks in the diagram below to label the steps of cross-validation.



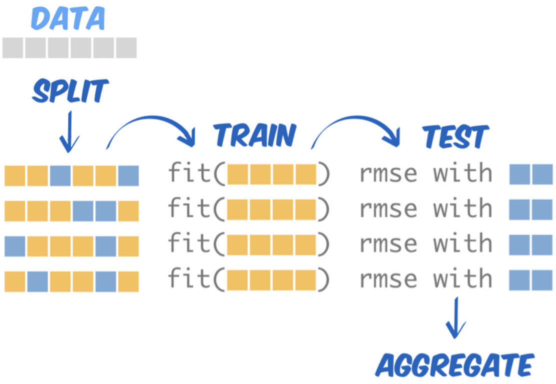
In your own words, answer:

2. Why should we split our data into training and testing sets?

(over)

3. Why does it make more sense to split our data into several different training and testing sets and average the results (i.e. to cross-validate) than to use a single training/testing set split?

4. The tidyverse code on the right is implementing a cross-validation strategy to evaluate a model with the ames data. Draw a line from each section of code on the right to the word in the diagram on the left that it is associated with. Can you tell what the code does?



```
1 ames %>%

2   vfold_cv(ames, v= 10, strata = Sale_Price) %>%

    mutate(
      train_set = map(splits, training),
      trained_model = map(train_set,
3       ~fit(Sale_Price ~ Gr_Liv_Area,
             model = lm_spec, data = .x)),

      test_set =  map(splits, testing),
      rmse = map2_dbl(trained_model, test_set,
4       ~rmse_vec(predict(.x, new_data = .y)$.pred,
                  .y$Sale_Price))
    ) %>%
    summarise(
5     mean = mean(rmse), sd = sd(rmse)
    )
```