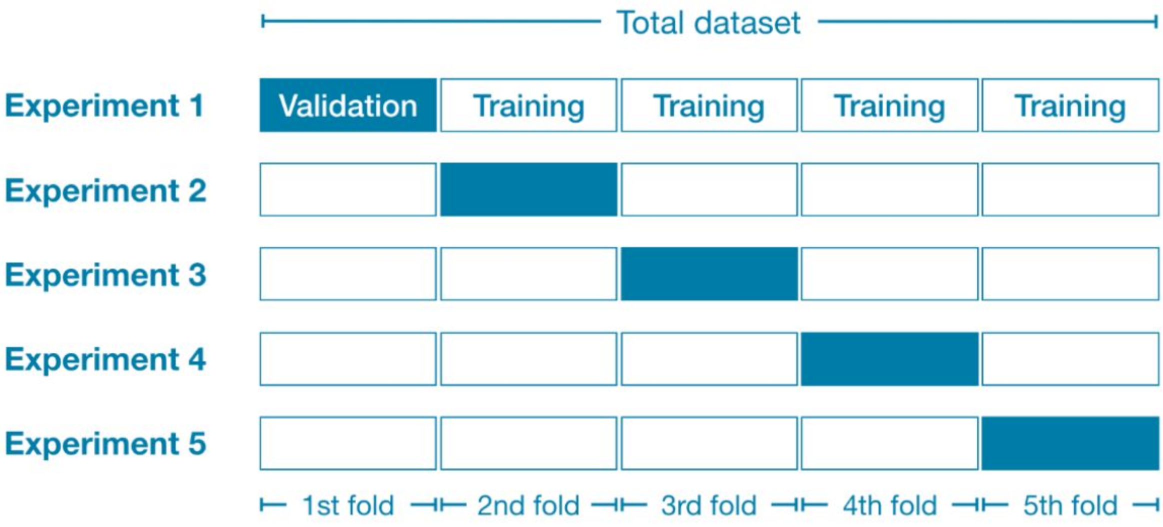
Reference: https://www.kaggle.com/alexisbcook/cross-validation

- Machine learning is an iterative process.
- Cross-validation used for better measures of model performance
- In general, the larger the validation set, the less randomness (aka "noise") there is in our measure of model quality, and the more reliable it will be. Unfortunately, we can only get a large validation set by removing rows from our training data, and smaller training datasets mean worse models!

How cross-validation work?



- Divide the training set to 5 folds ; 20% of the full data set
- Run 5 experiments, each time we each fold as a validation or hold out set and everything else as training data.
- Putting this together, 100% of the data is used as holdout at some point, and we end up with a measure of model quality that is based on all of the rows in the dataset (even if we don't use all rows simultaneously).

When we should use cross-validation

- *For small datasets*, where extra computational burden isn't a big deal, you should run cross-validation.
- *For larger datasets*, a single validation set is sufficient. Your code will run faster, and you may have enough data that there's little need to re-use some of it for holdout.
- you can run cross-validation and see if the scores for each experiment seem close. If each experiment yields the same results, a single validation set is probably sufficient.