# Resampling Notes

Get more information from data.

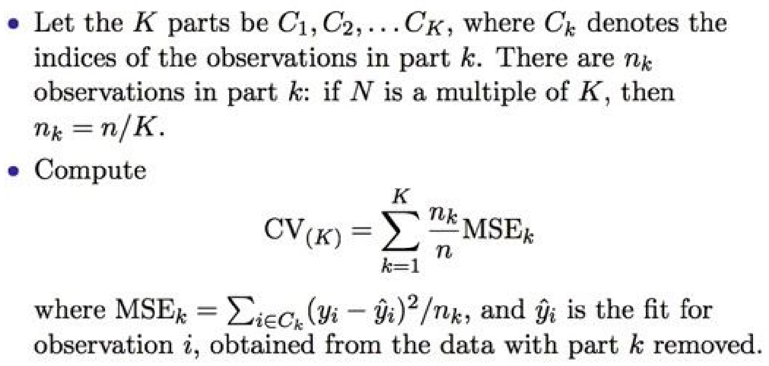
## Strategies:

### Cross Validation

1. Gives information about  test set error
2. Test error is how well we'd do on future data : "Test error is the average error we get while predicting the response of our statistical learning model on a new observation"
3. Test error calculated by holding out a dataset.
4. Training error can be very different from the test error.
5. Test error increases with high variance, high complexity and low bias in the dataset. (probably overfitting)
6. Data division = Training set and validation/ hold out set (used to test model responses)
7. If the data set is small just division might be wasteful as
8. Drawbacks
    a. Very variable splits
    b. Wasteful as only a subset can be used to train
    c. May lead to overestimation of test error

### K Fold Validation

1. Overcomes some drawbacks of cross validation
2. Widely used
3. Random division into k equal parts (often 5 or 10)
4. Use k-1 to train and the last one to test
5. Repeat with different forms
6. Combine results
7. Special case is leave one out cross validation



- Let the $K$ parts be $C_1, C_2, \ldots C_K$, where $C_k$ denotes the indices of the observations in part $k$. There are $n_k$ observations in part $k$: if $N$ is a multiple of $K$, then $n_k = n/K$.
- Compute

$$\mathrm{CV}_{(K)} = \sum_{k=1}^{K} \frac{n_k}{n} \mathrm{MSE}_k$$

where $\mathrm{MSE}_k = \sum_{i \in C_k} (y_i - \hat{y}_i)^2 / n_k$, and $\hat{y}_i$ is the fit for observation $i$, obtained from the data with part $k$ removed.

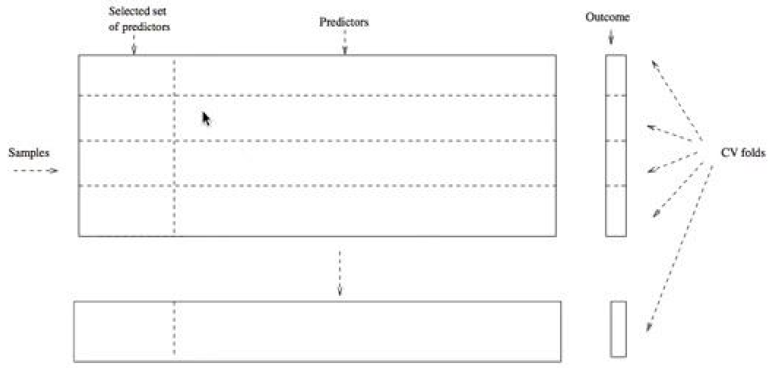
## How can you go wrong while cross validating?

**Wrong way:**

If we have a data set with more features than data points (example genomics data or healthcare data) then filtering out data based on the top predictors and then applying cross validation to divide dataset. This causes a huge bias as in the filtering step all the outcome variables have been viewed and filtering has been done based on that.

**Right way:**

First divide dataset and then filter. Right way:



# Bootstrap:

1. Good for understanding standard deviation and variability of dataset. Also gives a
2. Bias detection
3. Sampling(uncorrelated data) with replacement
4. Basically if you sample enough times then the standard deviation in the dataset gives you the error.
5. There is some overlap about 2/3rd data appears in bootstrap samples.

# Alternatives:

Mathematical strategies that estimate test error from training error. Cp statistic, AIC, BIC