During this class, we learn some techniques to detect problems, split our known data into several parts with applications, and the ways to distinguish the most effective model with several indicators.

From my point of view, this class's most valuable concept is knowing the difference between validation and test set. The validation set is defined to tune the hypo-parameter inside a specific model. For example, to recognise the highest power in multinomial cases or to find the minimum point in a separate tree’s end. Also, from a more general view, to change the loss from MSE to MAE or Friedman MSE counts (if the cost is above/ below the limit is different, define the specific loss). Doing research and combining different methods may improve the results. While the test set is selected from different models, each of them has its best parameter. Cross-validation is the method of splitting the data into folds, taking each fold as the test set separately to evaluate model results, which makes the result more convincing.

And then comes to the penalty part. LASSO adds a term with two-norm of the parameters and gets a sharper penalty, while Ridge adds the absolute value and pays attention to shrinking the absolute value of parameters. In a real case, we define the extreme penalty we want after normalisation, and then the LASSO will put some coefficients to zero. From the remaining part, we can see the more valuable factors, which are more in reducing the complexity.

Last, a part should be mentioned is the real efficiency of the final model. Many conditions may result in the less efficiency of the model. Such as the dataset size is not big enough, we use all past data to predict the future (which is unknown), or the model itself is meaningless.