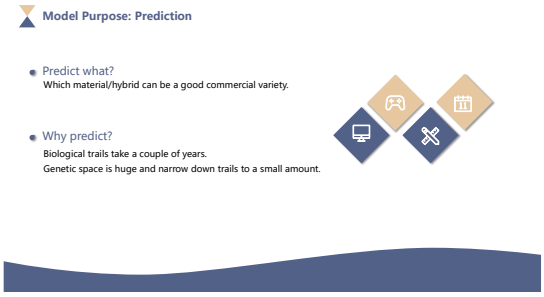
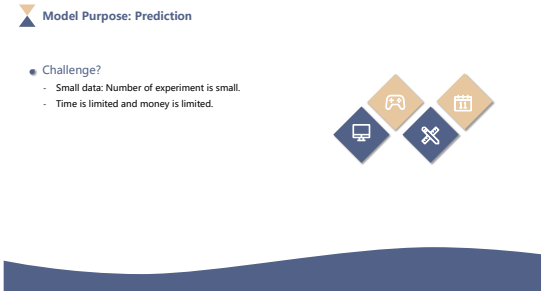
Presentation Notes

Ok, for the next part, I will introduce the model part in our interview. My teammates and I all agree with that as a statistician or a data scientist, model is the most important part, so we asked some questions about model.

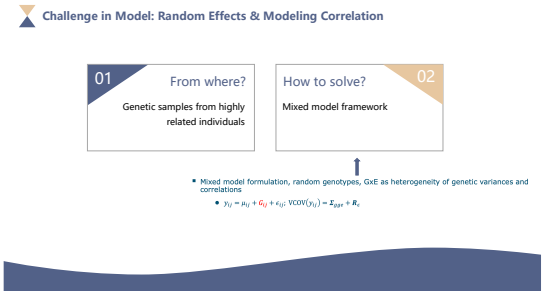


We totally proposed two main questions. The first one is about model purpose. We all know that we can use a statistical model to solve kinds of problems, like predicting future data, classifying objects, or just finding effective factors. So, we asked Dr. Marcos what is the main purpose of models in his work. The answer is prediction. Since Dr. Marco often works with samples from biology field and genetics field, so what he needs to predict is which material or which plant or which hybird / ˈhaɪbrɪd / will be a good commercial variety in the future. And he also told us reasons. Firstly, one biological trail can take long time, like 8 years or 10 years. The time cost is really high. Secondly, the genetic space is too huge. If we want, we can trail millions of combinations from genes, but that's impossible. So, we need to find the most valuable trails and do them. That's why we need to predict.

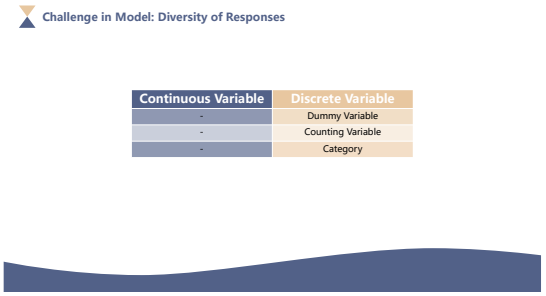


And Dr. Marco also mentioned that the challenge he met in prediction is small data problem. Here, small data problem means the number of trails we can do is small,

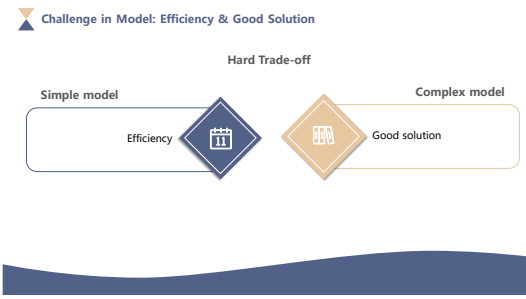
thus the train data we can use to predict is small. The reason for small data is also limited time and limited cost.



In the second question, we asked Dr. Marco which part is most complex or most difficult in his models. He gave us three challenges that he met. Now let me introduce them one by one. The first challenge is that he needs to include random effects and modeling correlation in his models. Because when he built model by data from genetic samples and population samples, the random effects always exist and some highly-related samples, like samples from parent test and child test, request to model the covariance or correlation. And the solution to this challenge is applying mixed model framework. Here is an expression for mixed model which we got from our field-trip in Wageningen University. If you're interested, you can search more information about this model.



The second challenge is there are all kinds of response variables in models. We'll meet continuous variables and discrete / dɪˈskriːt / variables. In discrete variables, we'll solve with dummy variables, like we predict survival or not survival, counting variables, for example we predict how many plants will be influenced by one disease and even category / ˈkætəgɔːri / variables, like we predict the color of fruit after experiment. It sounds not very difficult to solve them, but we can image that each kind of response variable needs some model adjustment or even a totally new model. So, work with different kinds, it's really time-consuming.



The final challenge is efficiency and good solution are both required to be achieved. We have learned from our model courses that if we use a simple model, the time of calculating will be short and the results will be easier to interpret. But the model maybe doesn't fit well. And if we change to a complex model, the fitting will be better but the time of calculating will be long and the results can be hard to explain. So, it's impossible to get efficiency and good solution in same time. What we can do is just trading off between them. So, it's always hard.

That's all of we get in model part from our interviewee and hope something in this part could be helpful to you. Thanks for your listening and let me invite my teammates Mikdad for the last part.