**Model Part Transcript**

1. Q: Which part do you think is the most complex or or most time consuming?

A: For complex what I would bring in is we're working with genetics and typically exploring populations and that brings automatically the concept of random effects in our models. That's the reason why mixed models are such an important part of our data analytics.

At the same time, we are working with different type of traits because makes models and continue straight goes very well along.

Then we can apply that, say, we can cooperate information on parent test, so relationships that means required of modeling correlations and covariation in the data.

That's clearly another reason why mixed models are there.

But we also need to account for other kind of responses which are not continuous, could be binomial and counting.

How many plants have been affected by these X uh.

Also, categories same certain properties of plants rather than putting it on a numeric scale, you put it in in categories, uh, from certain original scale, if we like it, but also requires thinking, OK, how to analyze this type of traits.

Uh, and also another efficiency part computationally.

In many times the results need to be in a short period of time and decision had to be made and very soon after the data has been provided.

So, data analysis had to be fast and there is where the methodology should be such that actually trace efficiency with the robustness of the method. We need to have a good analysis for making good decisions.

But we don't really need to go to the ultimate more elaborate because we simply don't have the time to do that.

But we need to find the right balance because too simple model is also not a solution.

I always say actually I don't want to wait for a whole day for some analysis to be done. There's far too much.

2. Q: What is the main purpose of statistical model in your research?

A: So, the main purpose is about prediction and prediction with small data.

Because what we're doing in breeding is create a new genetic, new combinations, and with a relatively small sample.

Try to find out which are the ones that in future are going to be potential new commercial varieties. The whole process for developing a new variety can take easily 8 or 10 years. But when you have a new potential product, a hybrid, for example, it happened already. A lot passed already. Then it goes into the field and this has to be tested and find out whether it's going to be a good material or not.

So, it's predicting in the end and we're not interested in the goodness of fit.

So, it's explaining what happened in that particular trial.

We want to from this trial make a good assessment what in future that material would do because we have to narrow down from hundreds to one or two, that will eventually be a good commercial variety and some prediction is very central and have as good prediction as possible, it's the main challenge.

3. Q: What do you mean by small data?

A: The number of experiments you can run is small. The per year is limited and you cannot spend 10 years trailing your material to find out this proper material because in 10 years maybe things changed already in terms of what is needed. So, you have to quickly come to the market with material that can be a good material commercially, but a period of time of trialing is not too long as in a couple of years.

And then there's a limited number of trials that you can run, and actually you can afford.

4. Q: I will be using mixed models. Could you give us more examples?

The reason why I mentioned mixed models is essentially what we're looking at is a population and there's a population of individuals that are highly related genetically. So, with the typically modeling performances by modeling genetic effects that are related, it's in challenge. Because the material is actually connected by pedigree, so pedigree information is important in the moment of analyzing the data and that all comes quite naturally in the mix model framework.

5. Q: What would you say the high dimensional data that you are working with?

A: The genetic space is huge and the number of combinations that you can get it's infinite. And you're able to just sample a very tiny part of it and that's the part you actually put in the field to test. Then in a few rounds of selection you have to decide which is good and which is not.