Questions Kircher et al (2014) *Nature Genetics* 46:310-315.

A general framework for estimating the relative pathogenicity of human genetic variants

1. What is the problem the authors were tempting to tackle?

2. What strategy was used to simulate sets of likely "deleterious" variants and sets of likely neutral variants? Why do you think this strategy worked?

3. The authors evaluate a series of "univariate" strategies for classifying variants as deleterious. What were the most effective univariate annotations? How were these annotations derived?

4. The authors describe a series of empirical assessments of their classifier. Which did you find most interesting? Why?

5. The authors say that non-sense variants had the highest C-scores on average. Why do think that is? They also say that missense variants near the start of a protein coding gene had higher C-scores than those near the end. Why do you think that is?

6. Can you think of other settings where ensemble approaches have been useful? What do these have in common?

7. How were missing values for input annotations handled when building the model?

8. How were categorical annotations handled when building the model?

9. The authors used the LIBOCAS library to implement their classifier. What are some of the advantages of using a library like LIBOCAS?

10. What struck you most about the paper?