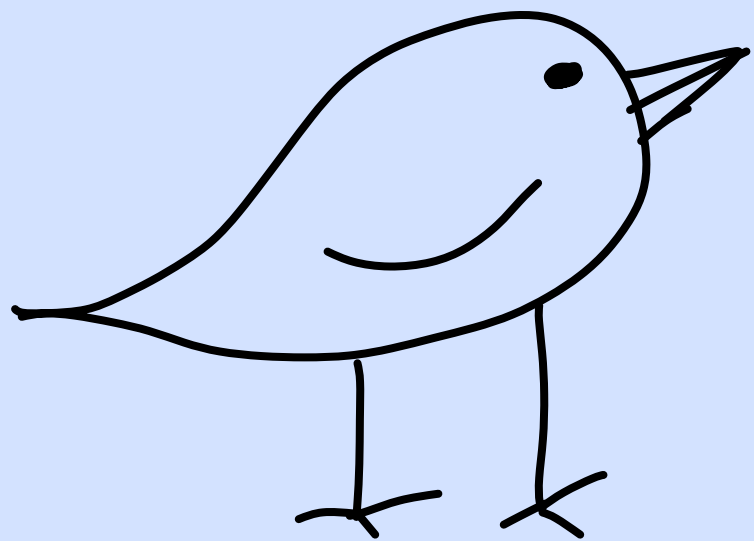


# Genomic Prediction

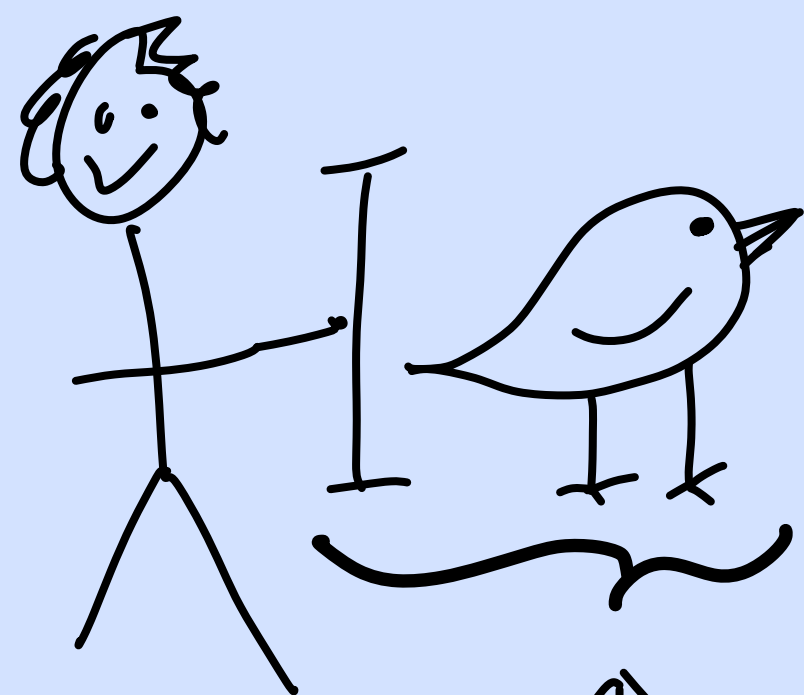
An application of gradient boosting



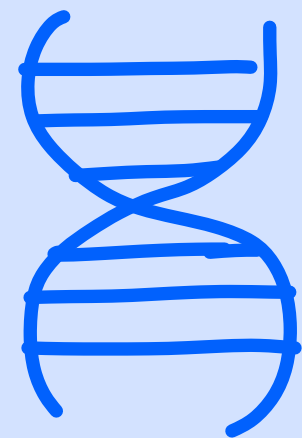
# Some Context

What is genomic prediction anyway?

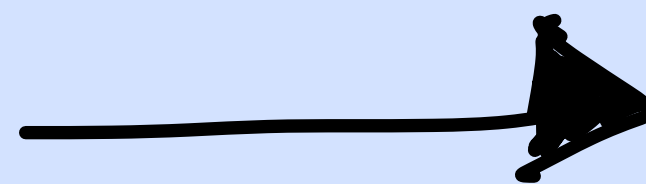
Biologist collects data of a trait



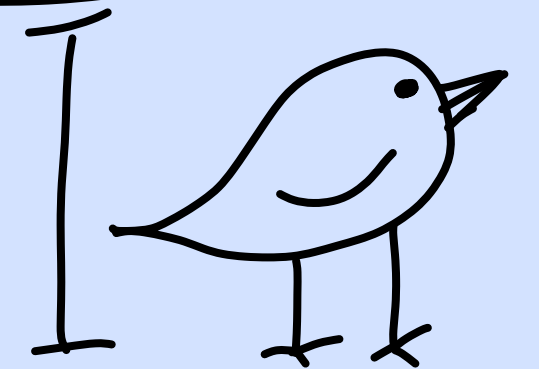
Trait



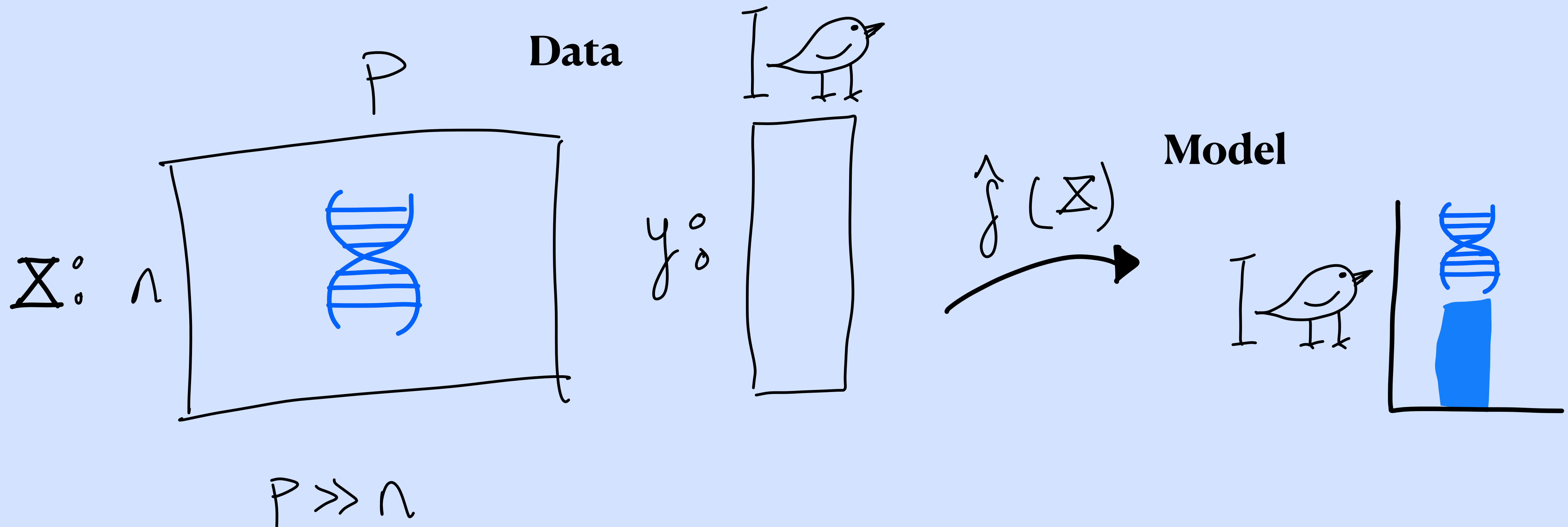
Genetics



How much of  
is heritable?



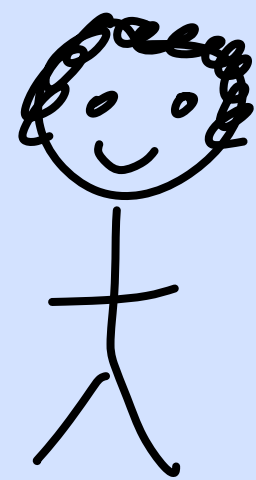
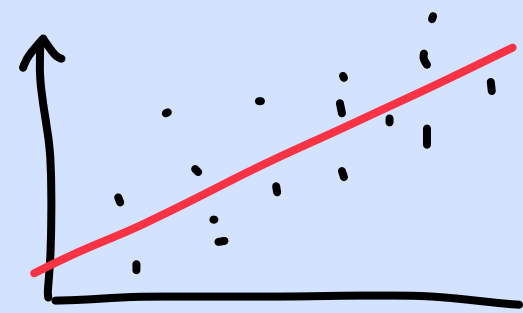
# The statistician's view of the problem



# How to build $\hat{f}(x)$ ?

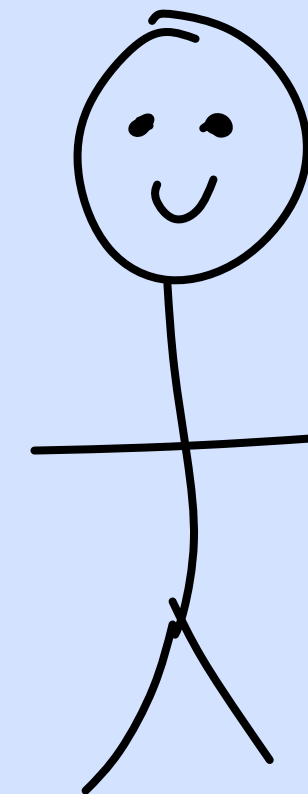
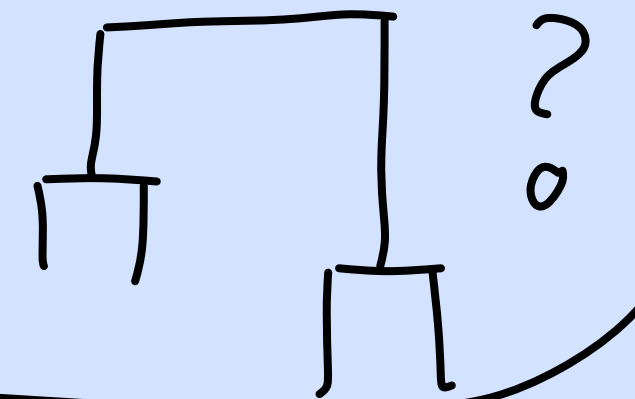
**Biologist:**

Traditionally we use



**Me who have taken statistical learning:**

Why not boosted



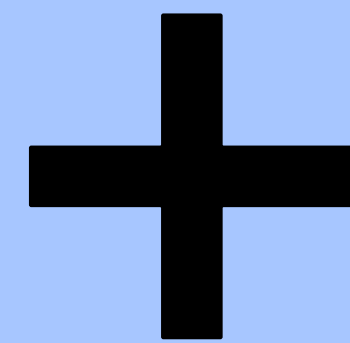
# How to build $\hat{f}(x)$ ?

10 Fold Cross Validation



CatBoost

*dmlc*  
***XGBoost***



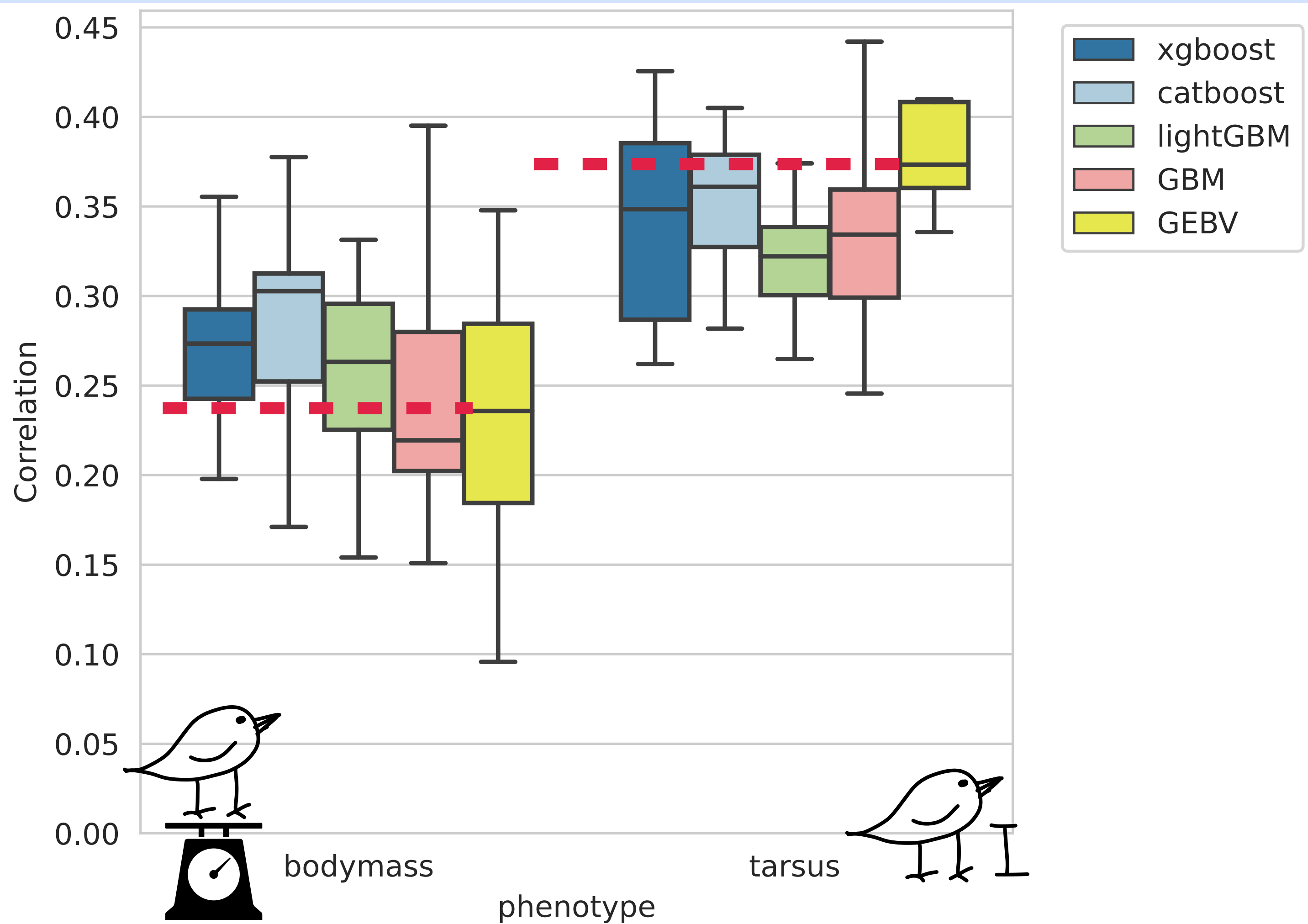
Hyperparameter Tuning



LightGBM

# Some Results

Higher is better



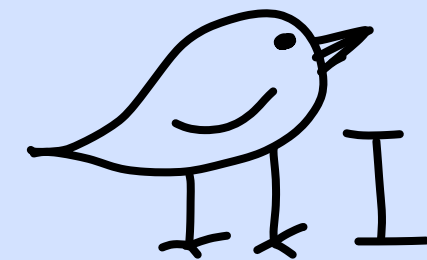
GBM: Basic Gradient  
Boosting

GEBV: Linear model

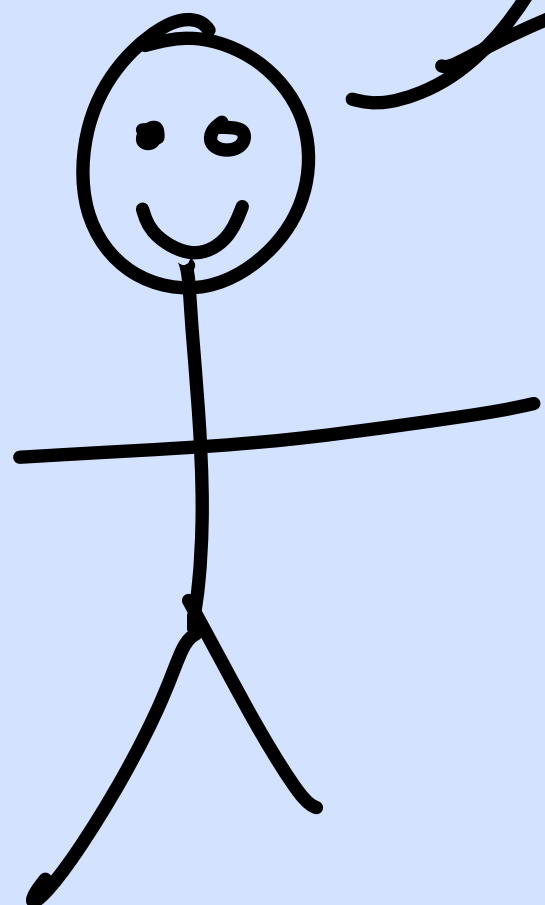
# Linear Booster

General Boosting Procedure:  $f_m(x) = f_{m-1}(x) + f(x)$  Weak learner

A linear model seems to be more suitable for

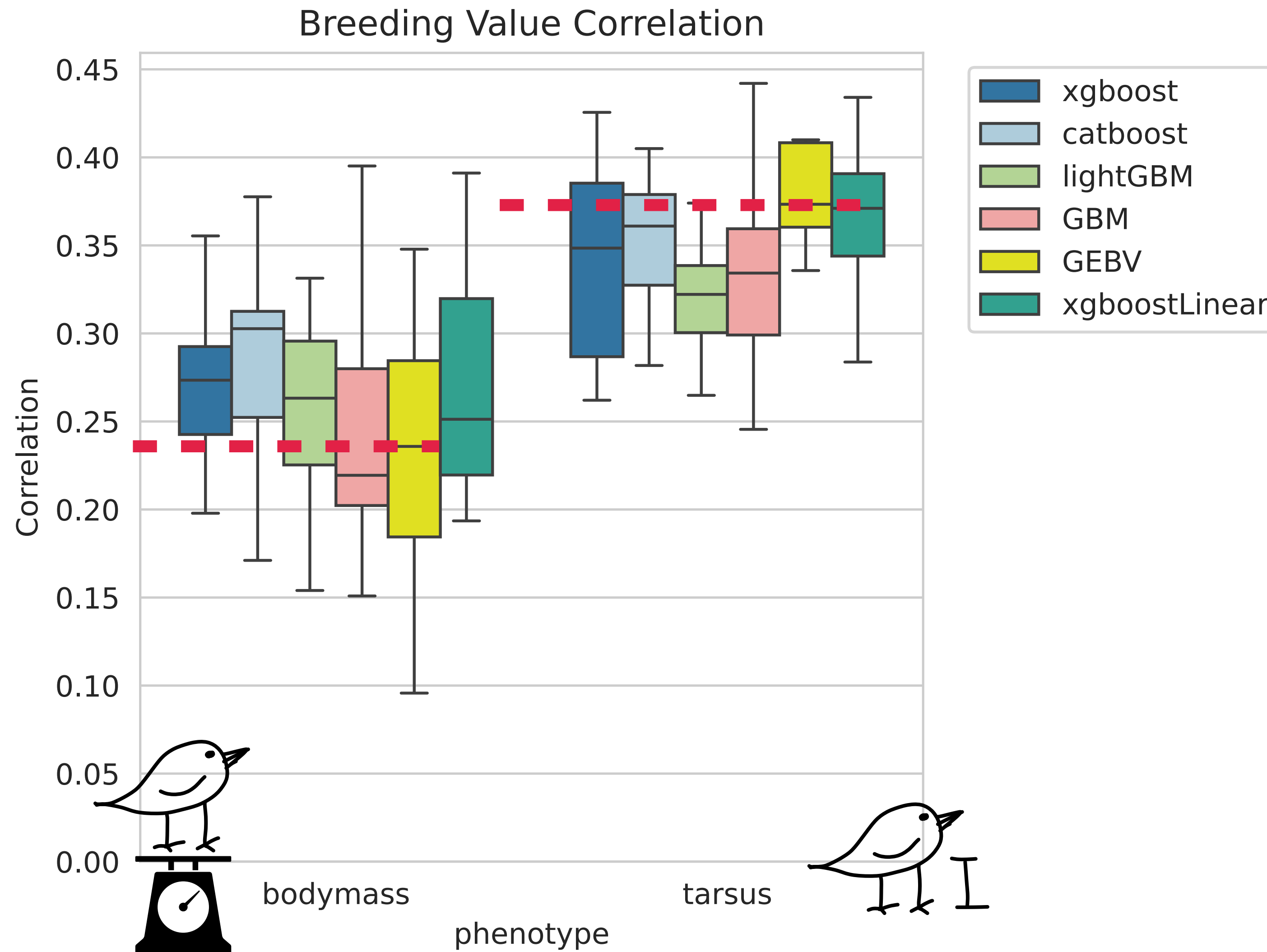


Why not use linear regression as weak learner in the boosting ?



# Linear Booster Results

Higher is better



GBM: Basic Gradient  
Boosting

GEBV: Linear model

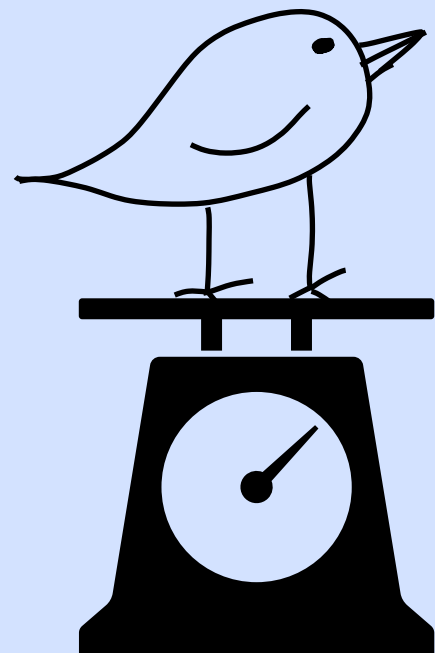


# Conclusion

Boosting is a real alternative to linear models in genomic prediction

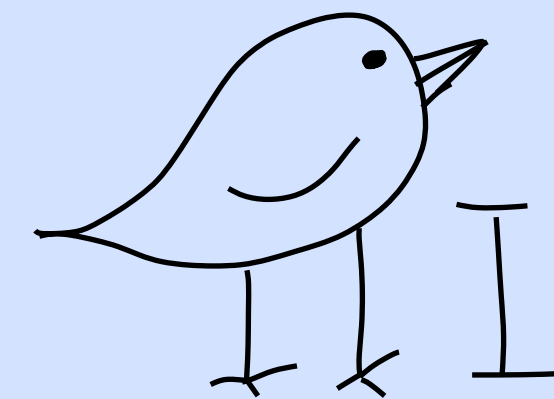
Body mass: Complex trait

⇒ Tree boosting works well



Tarsus: Less complex trait

⇒ Linear model more suitable



The underlying structure of the data matters the most!