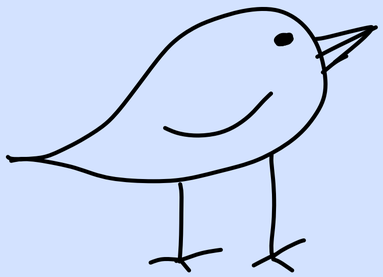
Genomic Prediction

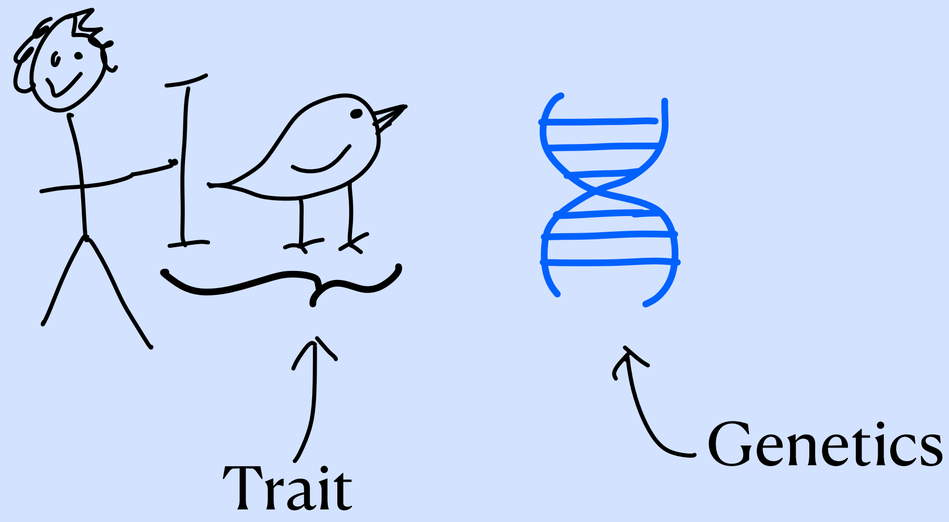
An application of gradient boosting

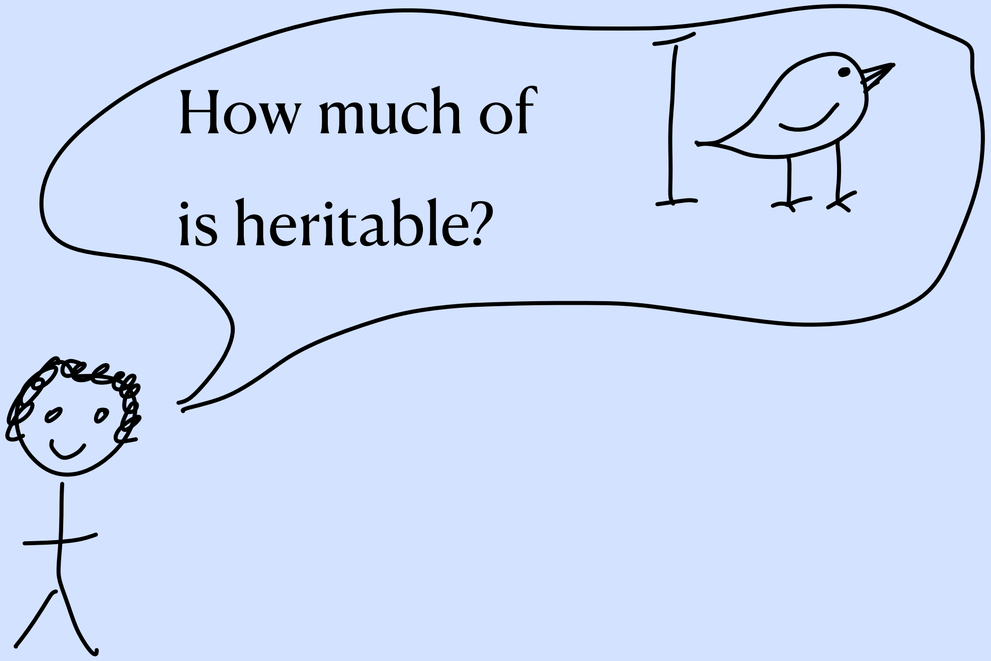


Some Context

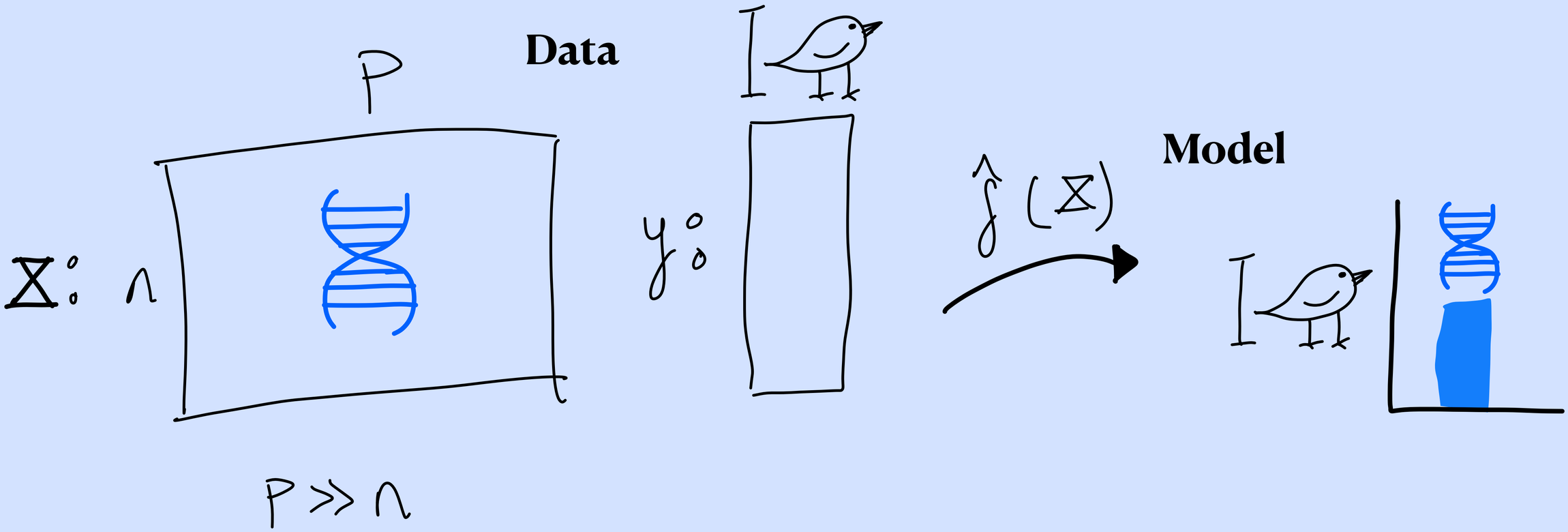
What is genomic prediction anyway?

Biologist collects data of a trait



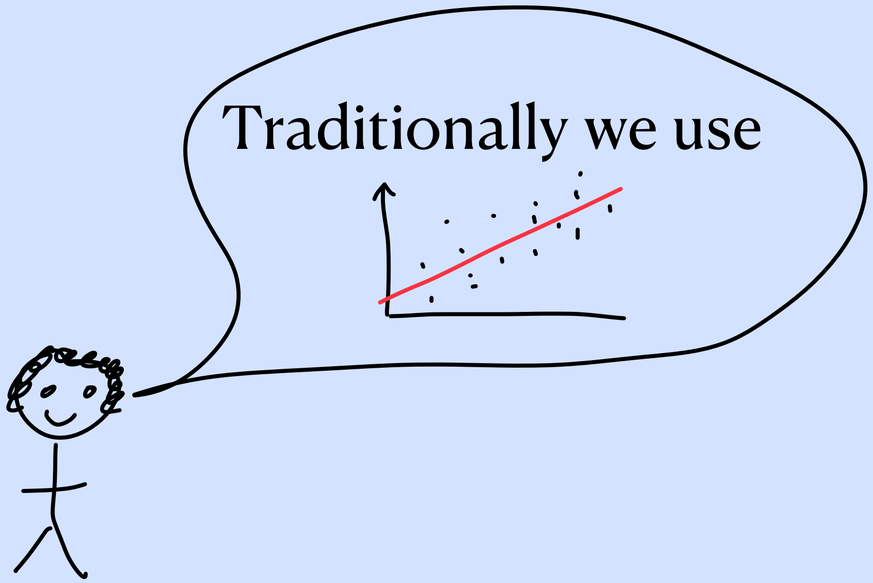


The statistician's view of the problem

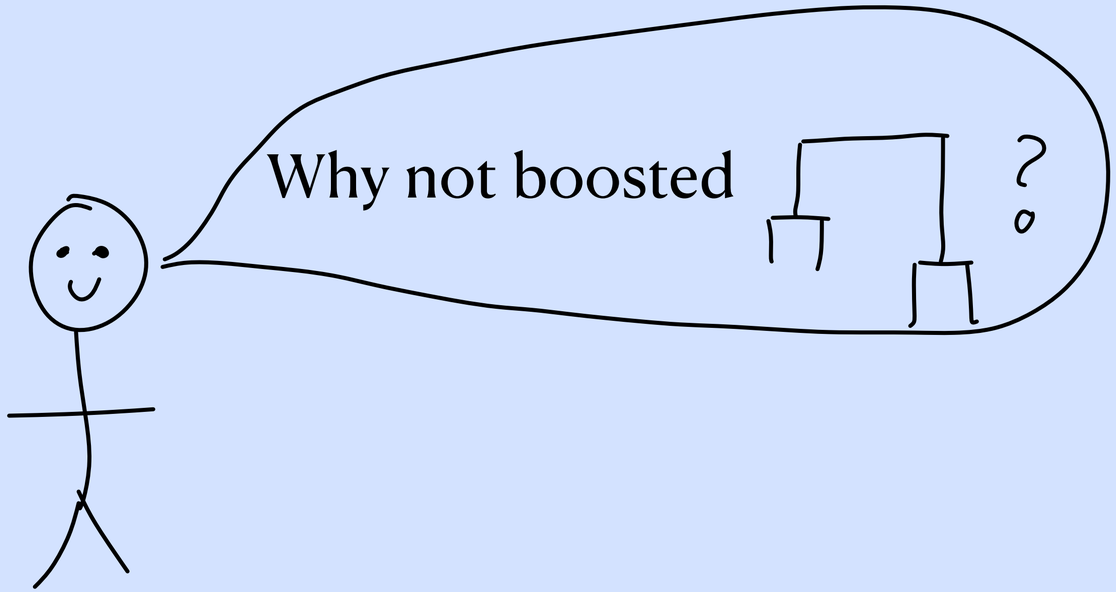


How to build $\hat{f}(x)$?

Biologist:



Me who have taken statistical learning:



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

10 Fold Cross Validation

CatBoost

dmlc XGBoost

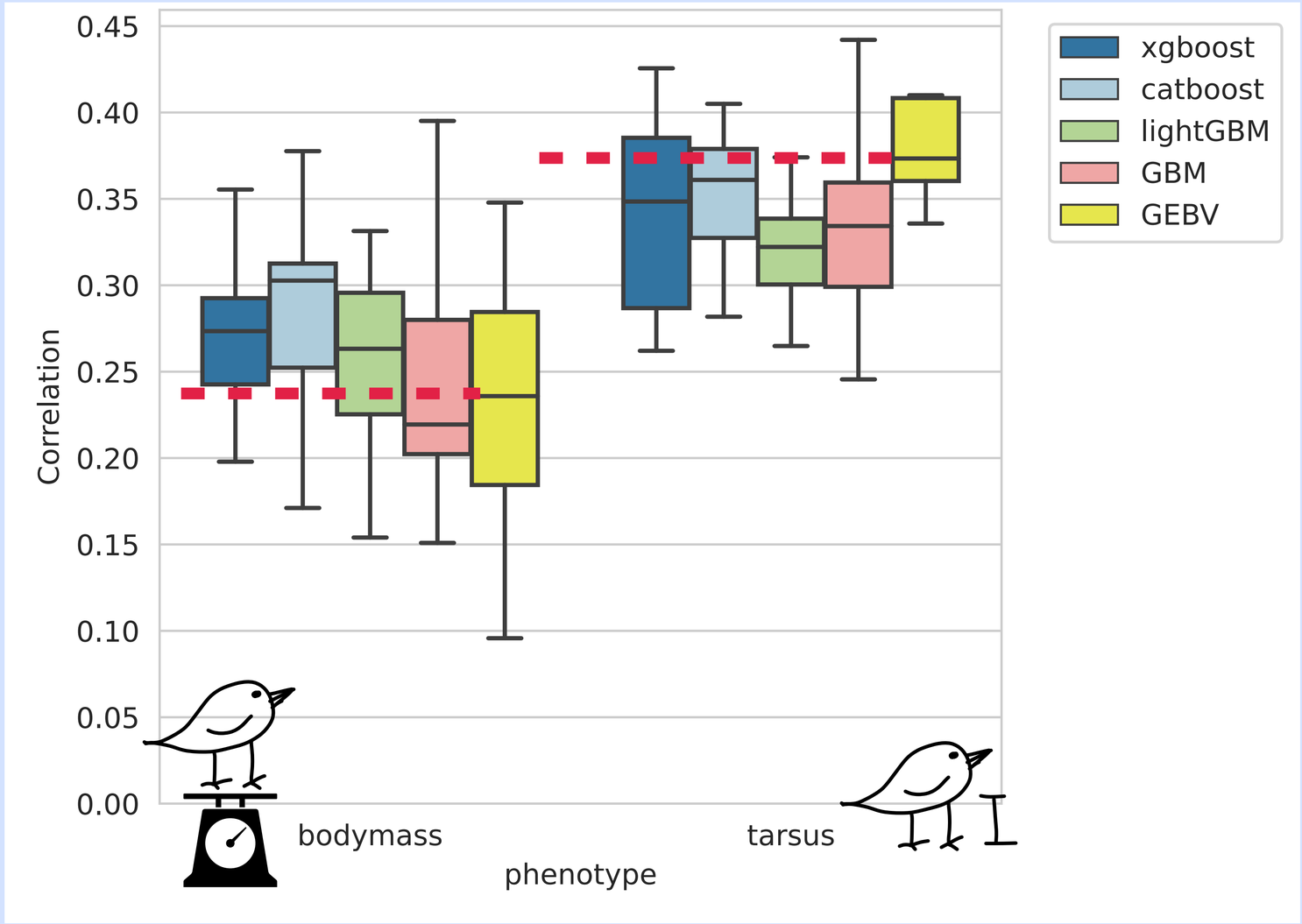
+

Hyperparameter Tuning

LightGBM

Some Results

Higher is better →



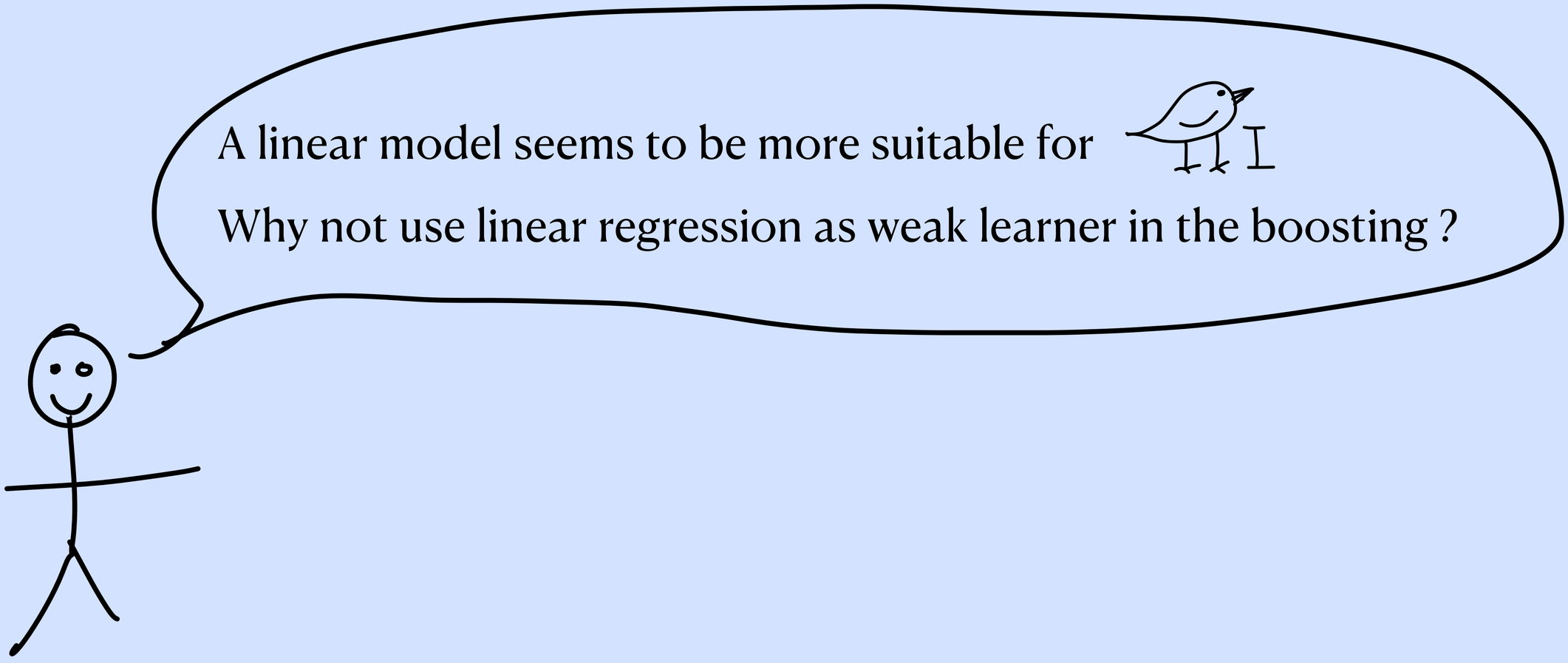
GBM: Basic Gradient Boosting

GEBV: Linear model

Linear Booster

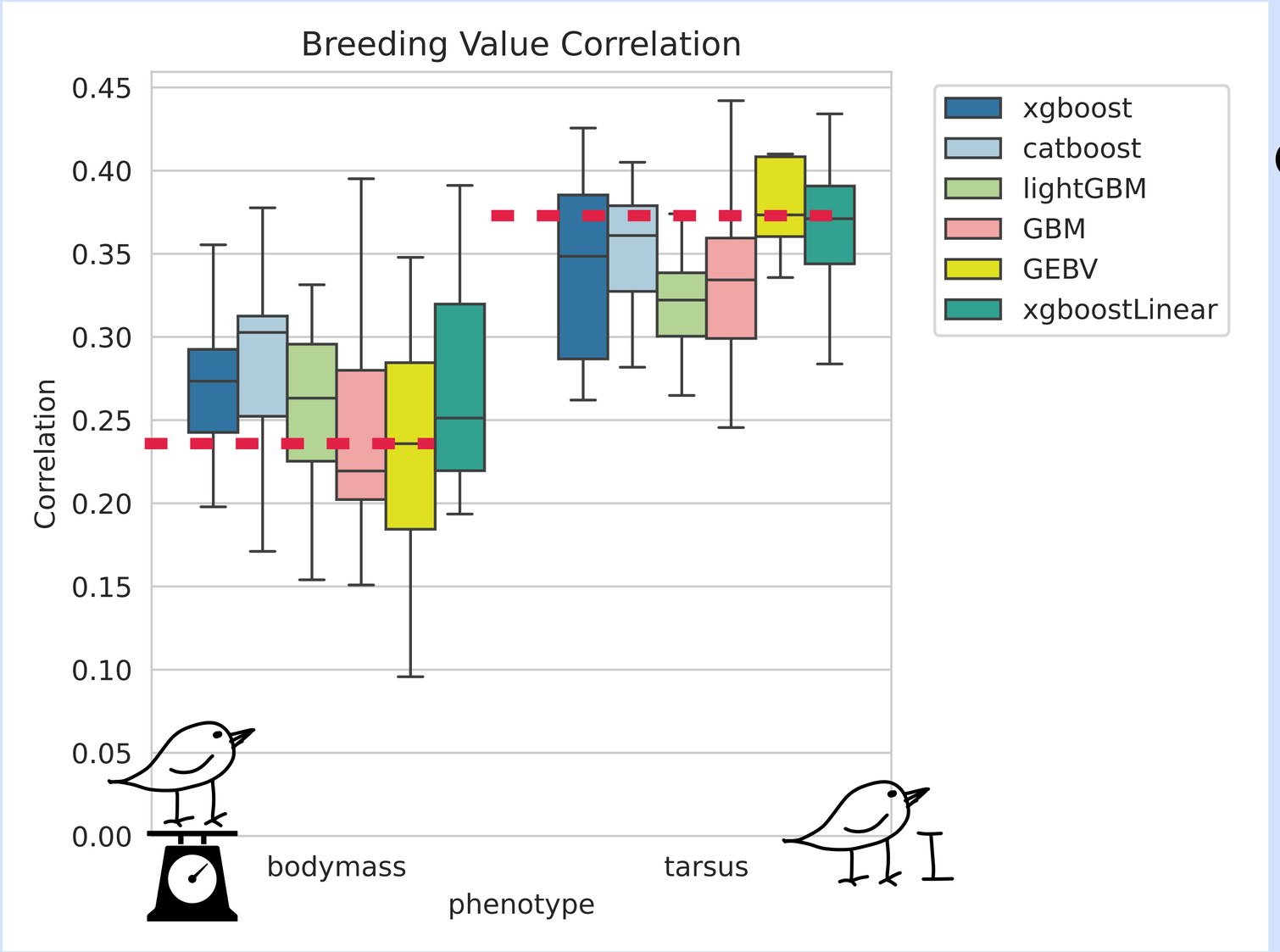
Weak learner

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



Linear Booster Results

Higher is better →



GBM: Basic Gradient Boosting

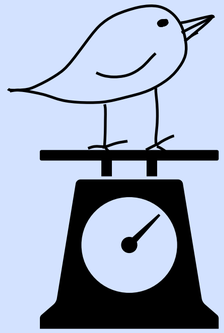
GEBV: Linear model

Conclusion

Boosting is a real alternative to linear models in genomic prediction

Body mass: Complex trait

$\Longrightarrow$ Tree boosting works well



Tarsus: Less complex trait

$\Longrightarrow$ Linear model more suitable

The underlying structure of the data matters the most!