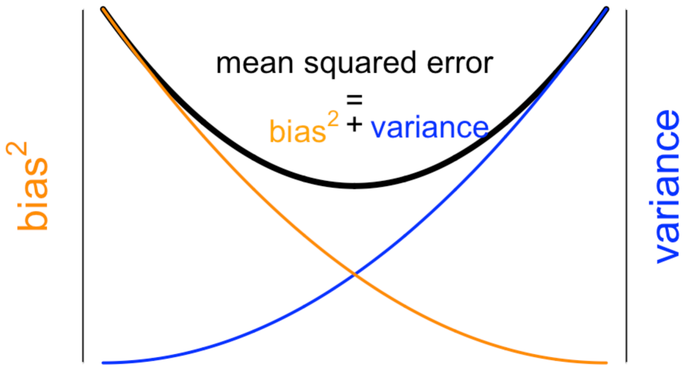
Lump/Split is an example of a GENERAL phenomenon in data science:

Some decision (model complexity, number of parameters, drilling down etc) triggers a tradeoff between **reliability** (e.g. low variance) and **validity** (e.g. low bias).

Mean Squared Error & the effect of model complexity

$$MSE(\hat{\theta}) = E\left((\hat{\theta}-\theta)^2\right) = \mathrm{var}(\hat{\theta}) + bias(\hat{\theta})^2$$



lump	←		→	split
simple	←	complexity	→	complex
few	←	#parameters	→	many
few	←	#components	→	many
large	←	penalty	→	small
heavy	←	prior weight	→	light

As you make a model more complex,

(adding parameters, splitting data into smaller subsets
with a different model for each group, ; etc)
it fits better, but eventually overfits.
It loses **reliability** (reproducibility)

- The model gains "degrees of freedom";
  so it ***can fit*** the data more closely.

- The data loses "degrees of freedom",
  so it ***can't critique*** the model as well.

- The effect size may get much bigger
  (if you chose the right split).

- The data in each subgroup is sparser
  so the variance is higher.

- We are closer to asking the right question for the individual…
  but with less accuracy as the sample size shrinks

- **Bias** is high when your study is asking the wrong question
  poor **Validity.**

- **Variance** is high when, on repeating the study,
  the estimates would change greatly:
  poor **Reliability.**