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

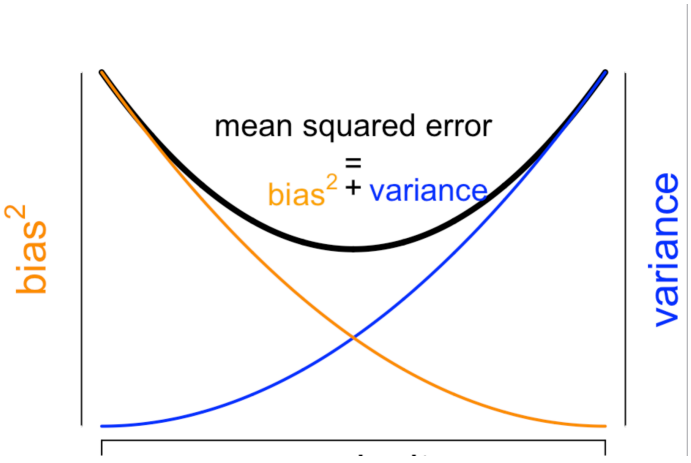
Some decision (model complexity, number of parameters, drilling down etc) triggers a tradeoff between reliability (e.g. low variance) and validity (e.g. low bias).

- As you make a model more complex and "free",
  it fits better, but eventually overfits.
  - 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.

- As you drill down into smaller subsets, leaning towards "splitting",.
  it fits better, but eventually overfits.
  - The effect size may get much bigger
    (if you chose the right split).
  - The data 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

Example: Mean Squared Error in regression; effect of model complexity

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



lump	←		→	split
simple	←	complexity	→	complex
few	←	#parameters	→	many
few	←	#components	→	many
large	←	penalty	→	small

heavy ← prior weight → light

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.; related to **Reliability.**

2.c) Example: hypothetical medical study demonstrating lumping versus splitting

The Problem: A new treatment is given to 100 patients. Of them, only 8 respond. But there is a subgroup of 5 in which 3 patients respond, yielding a response rate of 60%! Should the treatment be recommended for people in the subgroup?

	Group D	**Group L**	TOTAL
Responder	**3**	**5**	8
Nonresponder	**2**	**90**	92
TOTAL	5	95	100

What if D and L are:

2 alleles of a gene known to affect this drug's pharmacodynamics

2 alleles of one gene out of a hundred known to affect this drug's pharmacodynamics

2 alleles of one gene out of a hundred thousand; nothing known

D = dark hair, L = light hair

D = dark hair, L = light hair; hair color is strongly tied to ethnicity…

which is strongly tied to a key enzyme

2.d) Consequences of splitting, good and bad:

i. Decreased bias
ii. Increased variance
    - due to smaller samples sizes
    - due to decreased variation in treatments delivered (if not randomized)
iii. Hopefully ***greatly*** increased effect sizes
iv. Risks of multiple testing

2.e) Optimizing the lump/split compromise

i) For internal validity

Internal validity: the answer is sufficiently correct to apply to new patients "similar" to those in this study.
It will keep on working well here, for patients like these, even if we don't know why.

ii) For external validity

External validity: the answer is sufficiently correct to apply even to new patients from a different sampling catchment
(age, location, ethnicity, socio-economic, …).

The science is well grounded enough to generalize.

iii) Techniques:
Meaningful Bayes priors, Bayesian networks, hierarchical models, empirical Bayes,