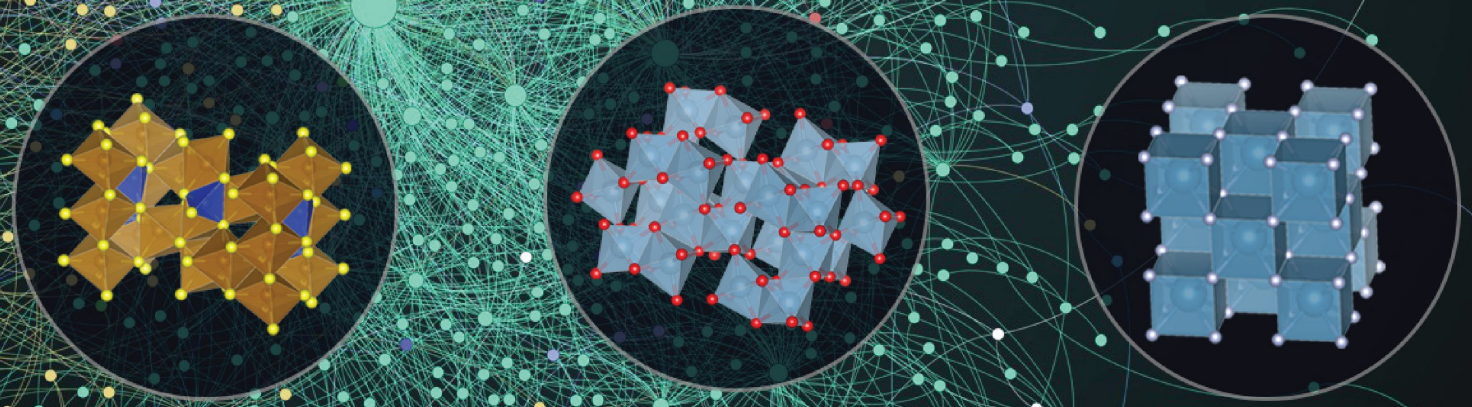
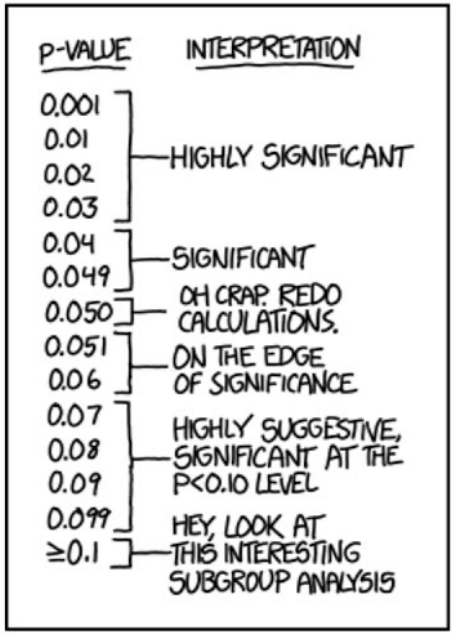
splitting data into test, train, validation



We need a test set to prevent p-hacking!

What is p-hacking?



https://imgs.xkcd.com/comics/p_values.png

Recall our linear models

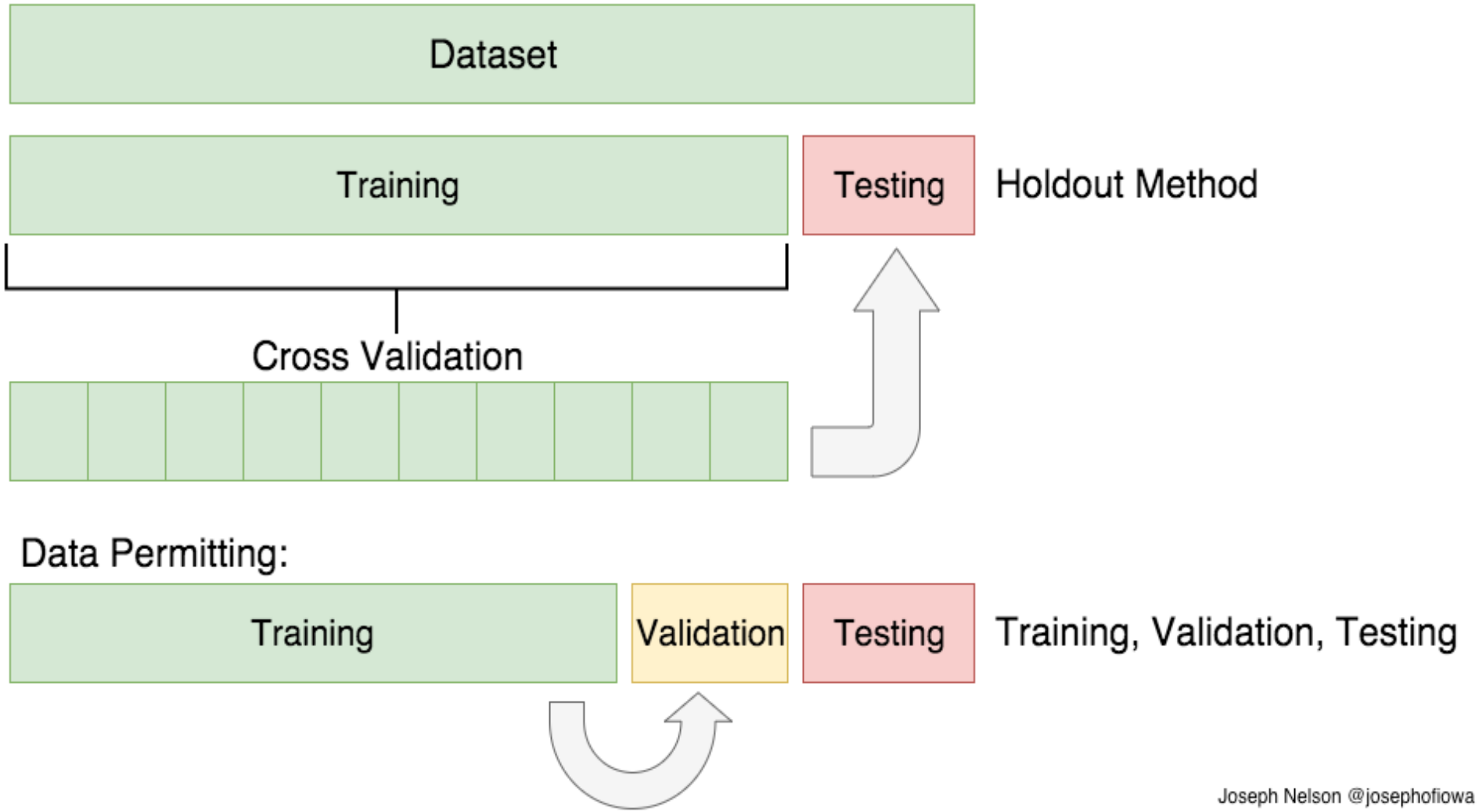
$$dependent\ variable = constant + parameter * IV + parameter * IV + \cdots$$

If a constant is zero, then that parameter doesn't contribute to the dependent variable

There are many ways to p-hack

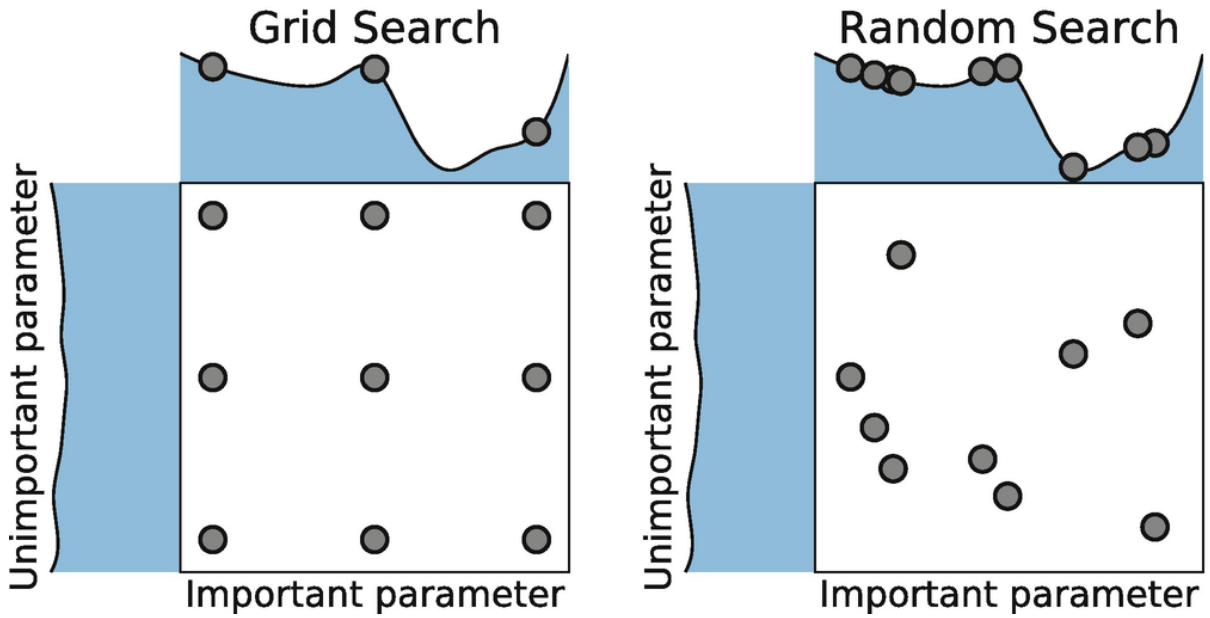
1. Stop collecting data once $p<.05$
2. Analyze many measures, but report only those with $p<.05$.
3. Collect and analyze many conditions, but only report those with $p<.05$.
4. Use covariates to get $p<.05$.
5. Exclude participants to get $p<.05$.
6. Transform the data to get $p<.05$.

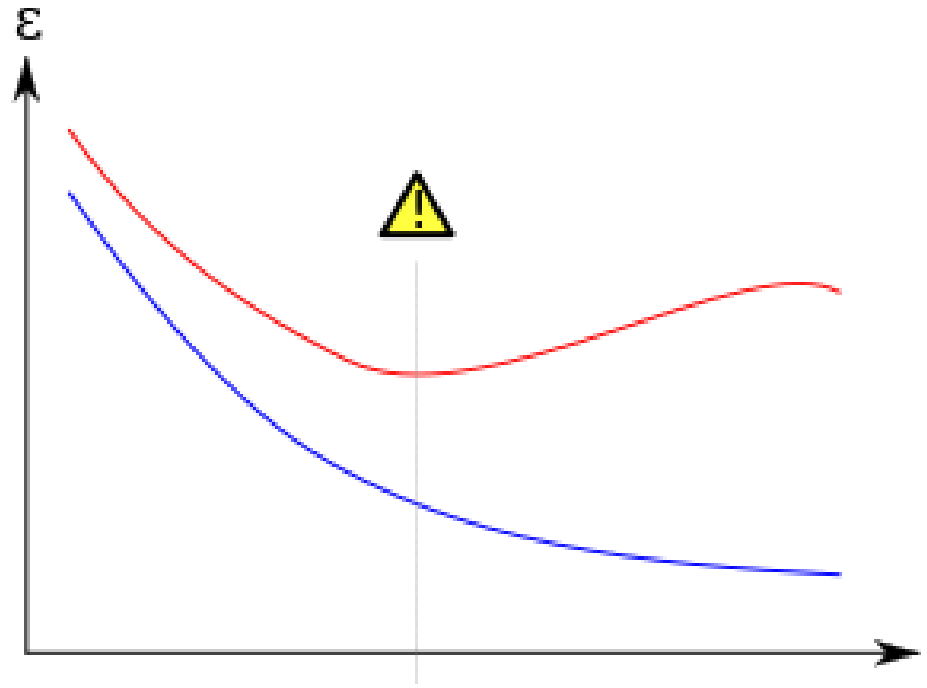
We prevent p-hacking by using distinct test/train/val datasets



Joseph Nelson @josephofiowa

Validation datasets allow us to tune the model hyperparameters

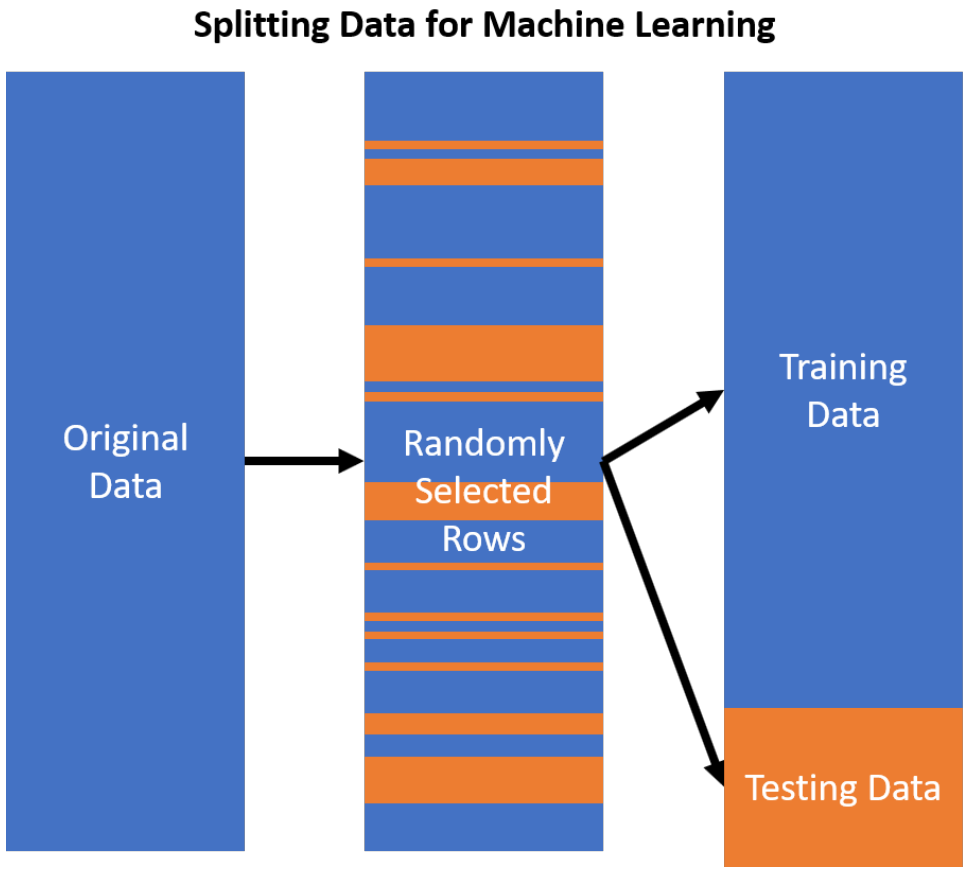


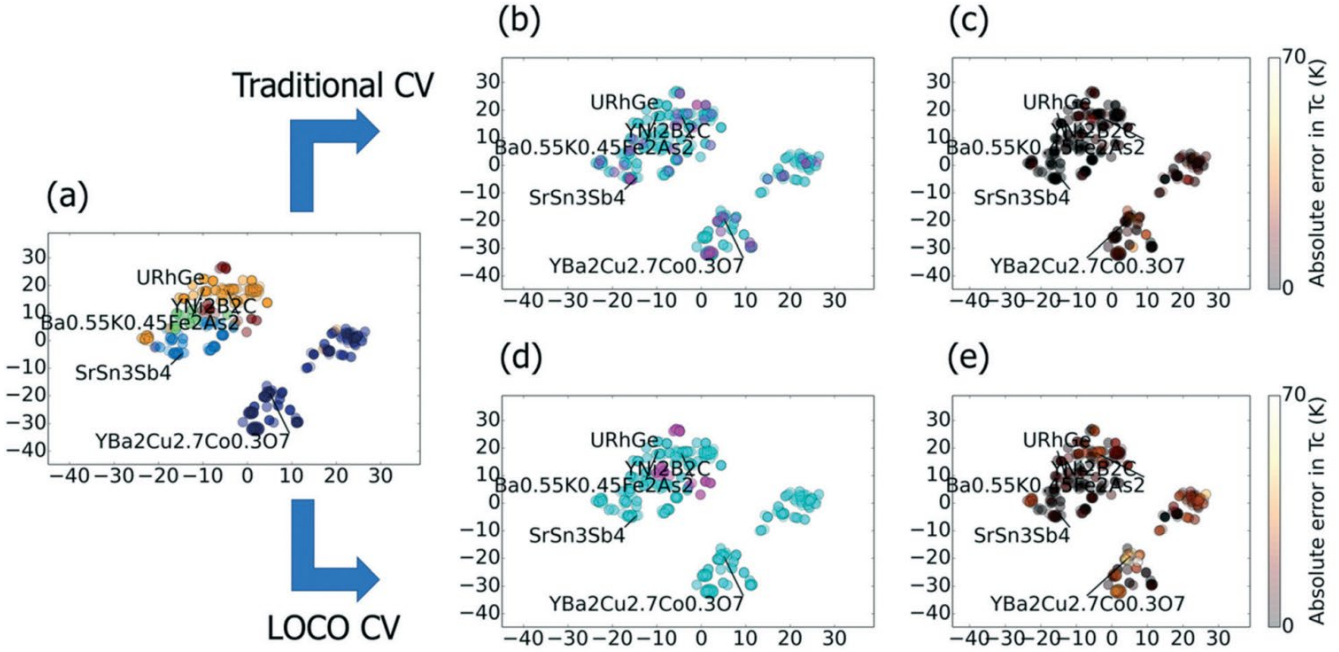


https://en.wikipedia.org/wiki/Overfitting

Are there right and wrong ways to split data??

Random vs Intentional





materials informatics best practices