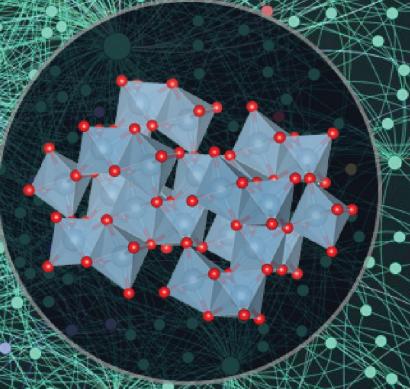
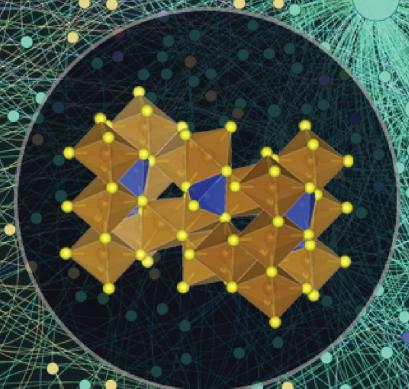


# splitting data into test, train, validation



We need a test set to prevent p-hacking

## What is p-hacking?

P-VALUE	INTERPRETATION
0.001	
0.01	
0.02	HIGHLY SIGNIFICANT
0.03	
0.04	SIGNIFICANT
0.049	OH CRAP, REDO CALCULATIONS.
0.050	
0.051	ON THE EDGE OF SIGNIFICANCE
0.06	
0.07	HIGHLY SUGGESTIVE,
0.08	SIGNIFICANT AT THE P<0.10 LEVEL
0.09	
0.099	HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS
≥0.1	

[https://imgs.xkcd.com/comics/p\\_values.png](https://imgs.xkcd.com/comics/p_values.png)



## Recall our linear models

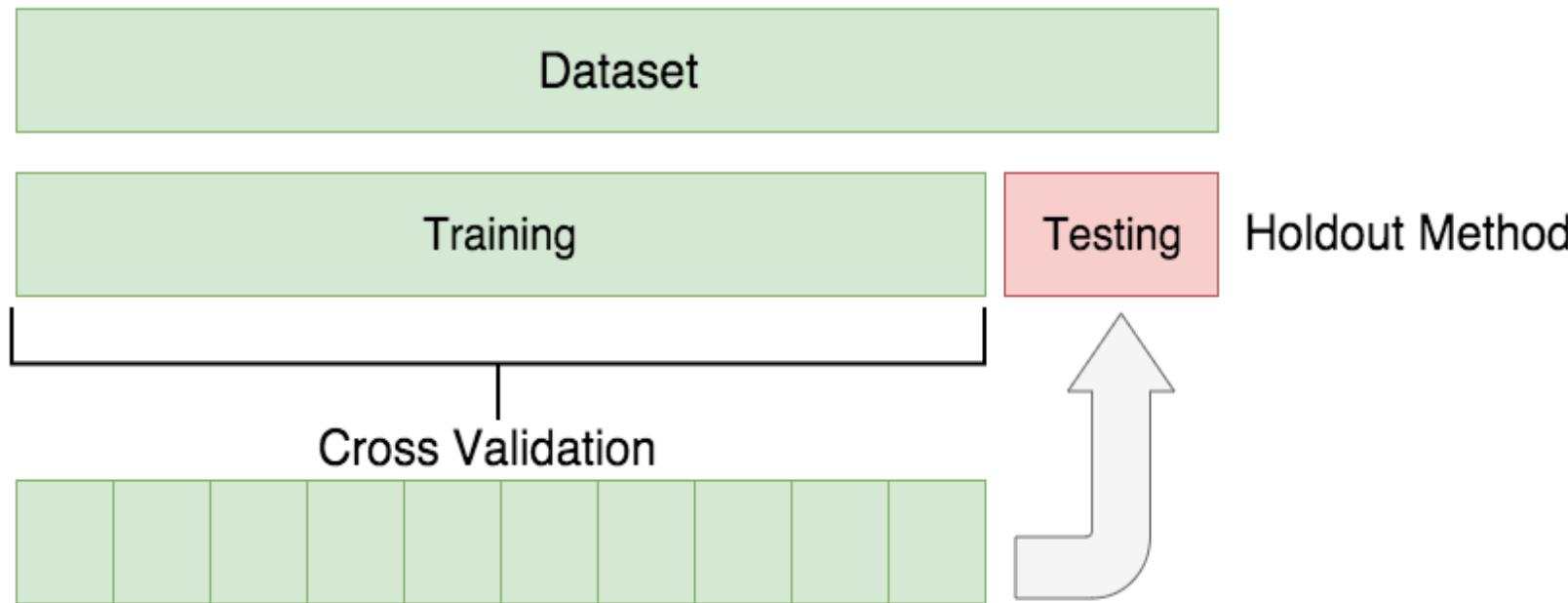
*dependent variable = constant + parameter \* IV + parameter \* IV + ...*

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

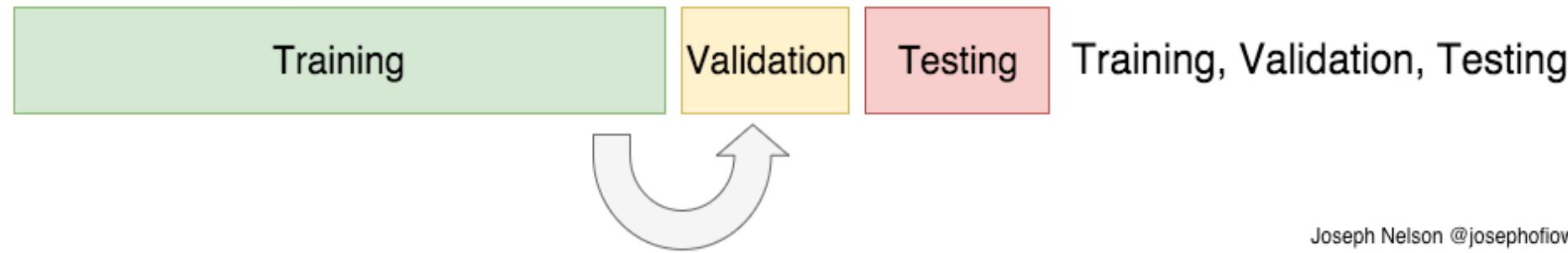
There are many ways to “p hack”

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 in machine learning through the use of different data sets

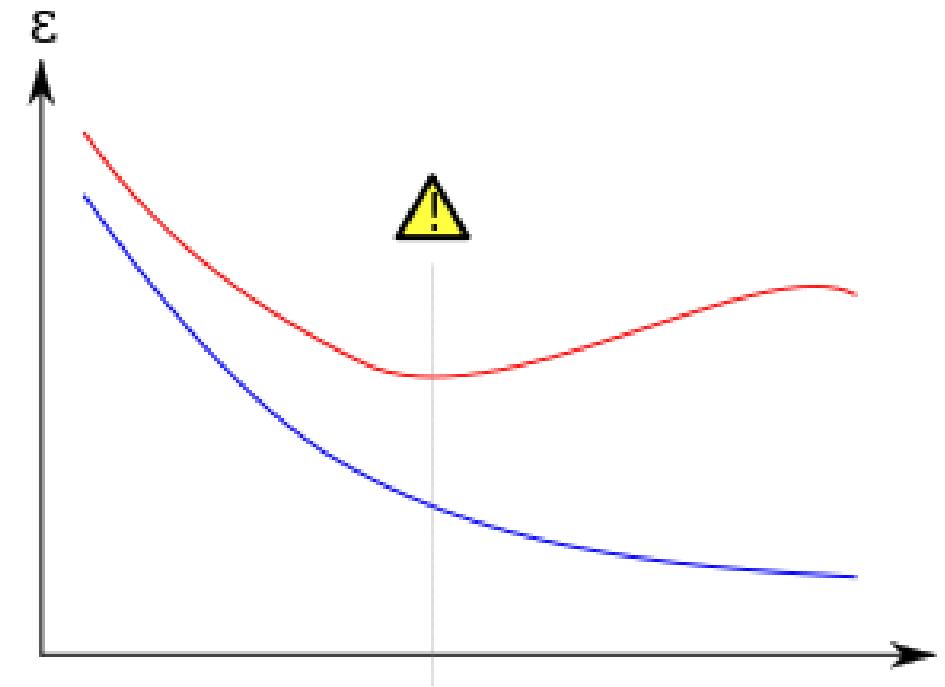
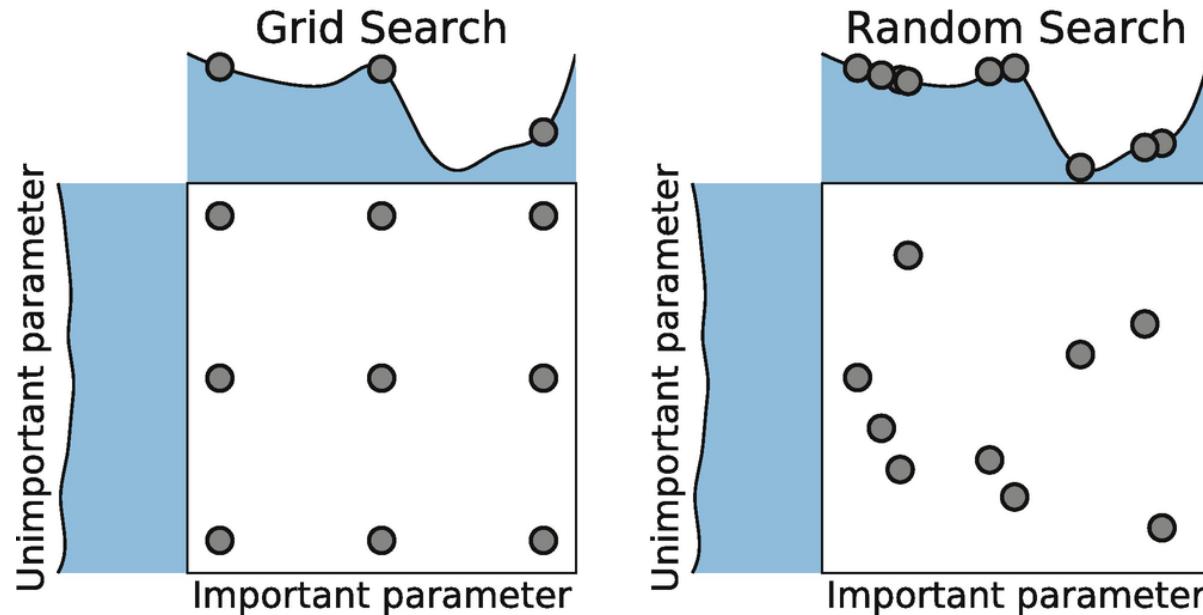


Data Permitting:

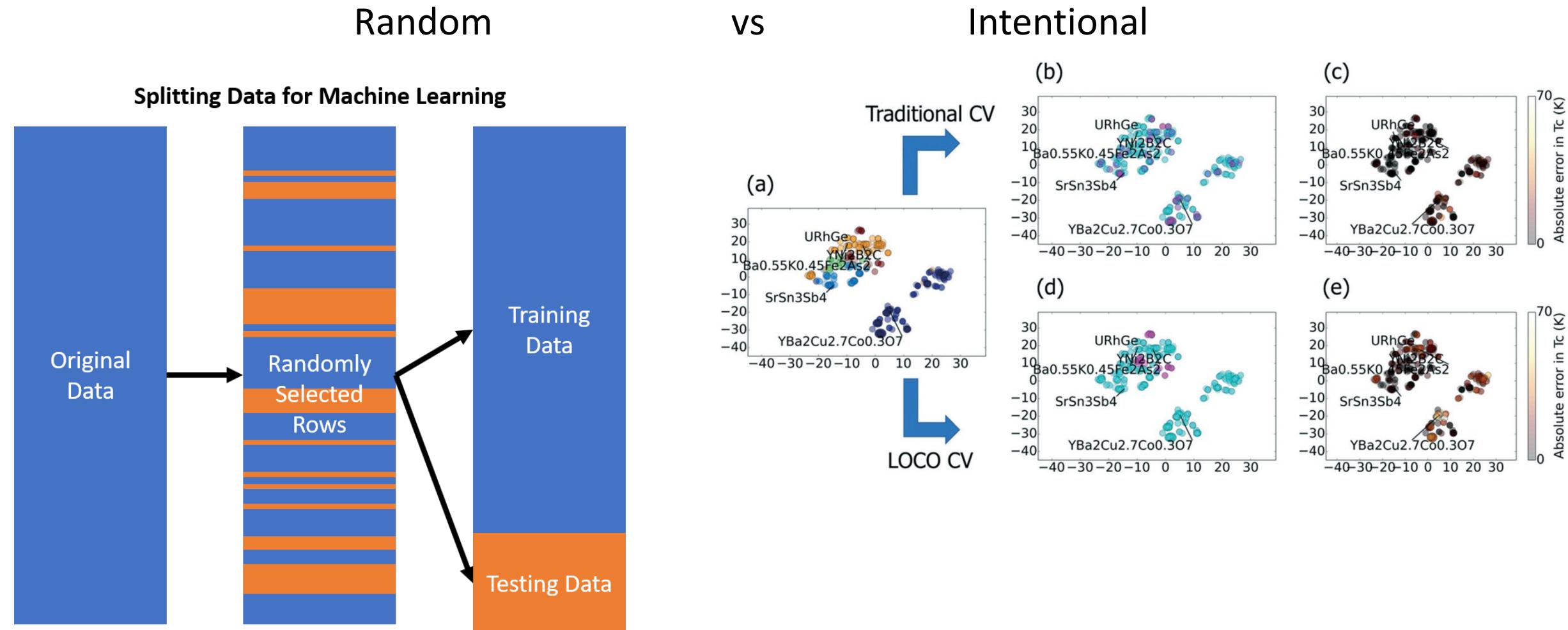


Joseph Nelson @josephofawa

Models are optimized on the training or validation data sets with care to avoid overfitting



# Is there a right way and a wrong way to split our data?



# materials informatics best practices

