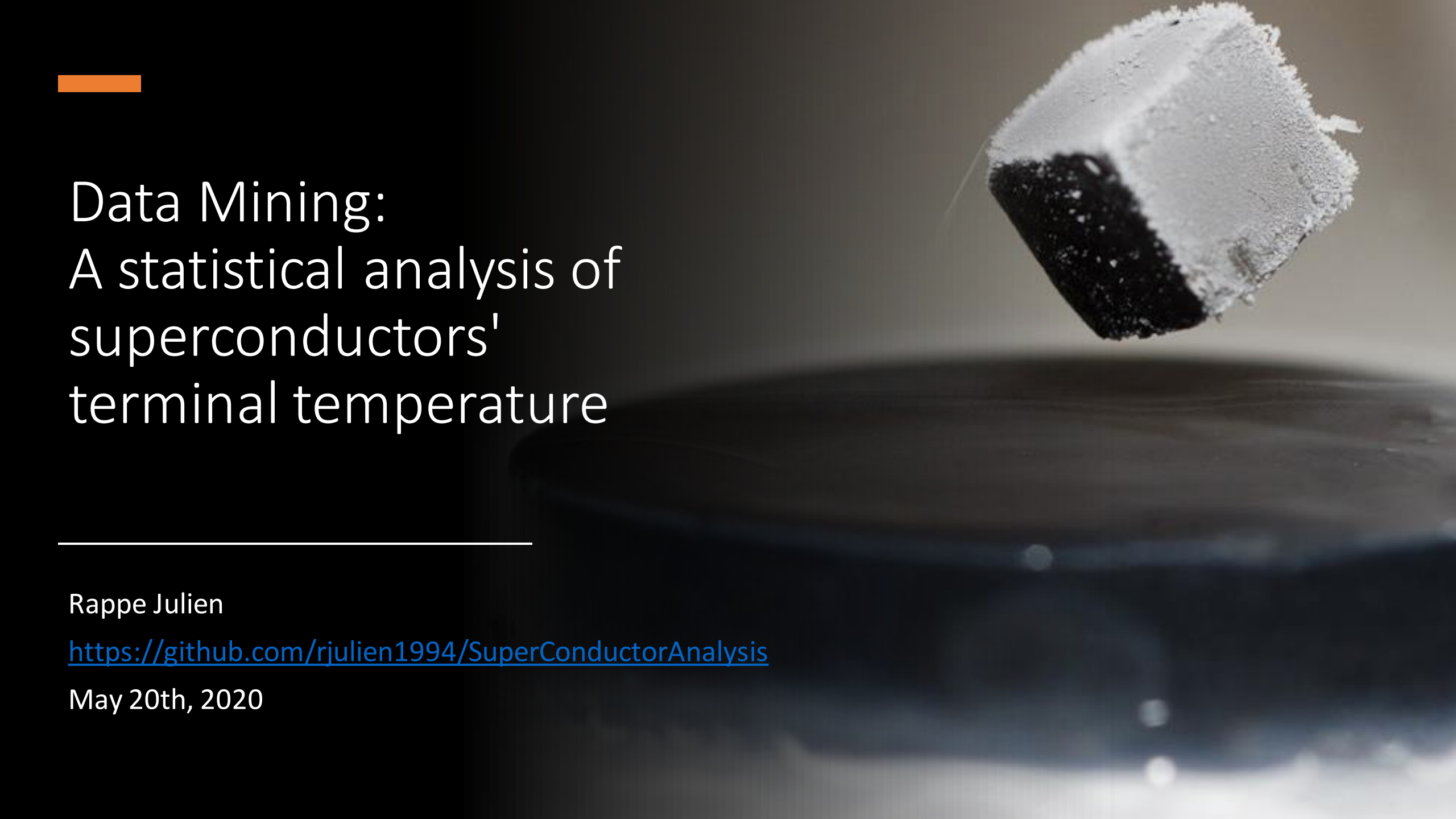


Data Mining:
A statistical analysis of superconductors' terminal temperature

Rappe Julien

https://github.com/rjulien1994/SuperConductorAnalysis

May 20th, 2020

Introducing the data

- **Objective:**

We want to estimate the critical temperature of different super-conductors based on their physical characteristics.

- **Data Set:**

The original data set had 82 variables and 21263 records with no missing data.

- **Predictors:**

The 81 attributes were collected in a lab environment and each record is already the result of averaging many experiments.

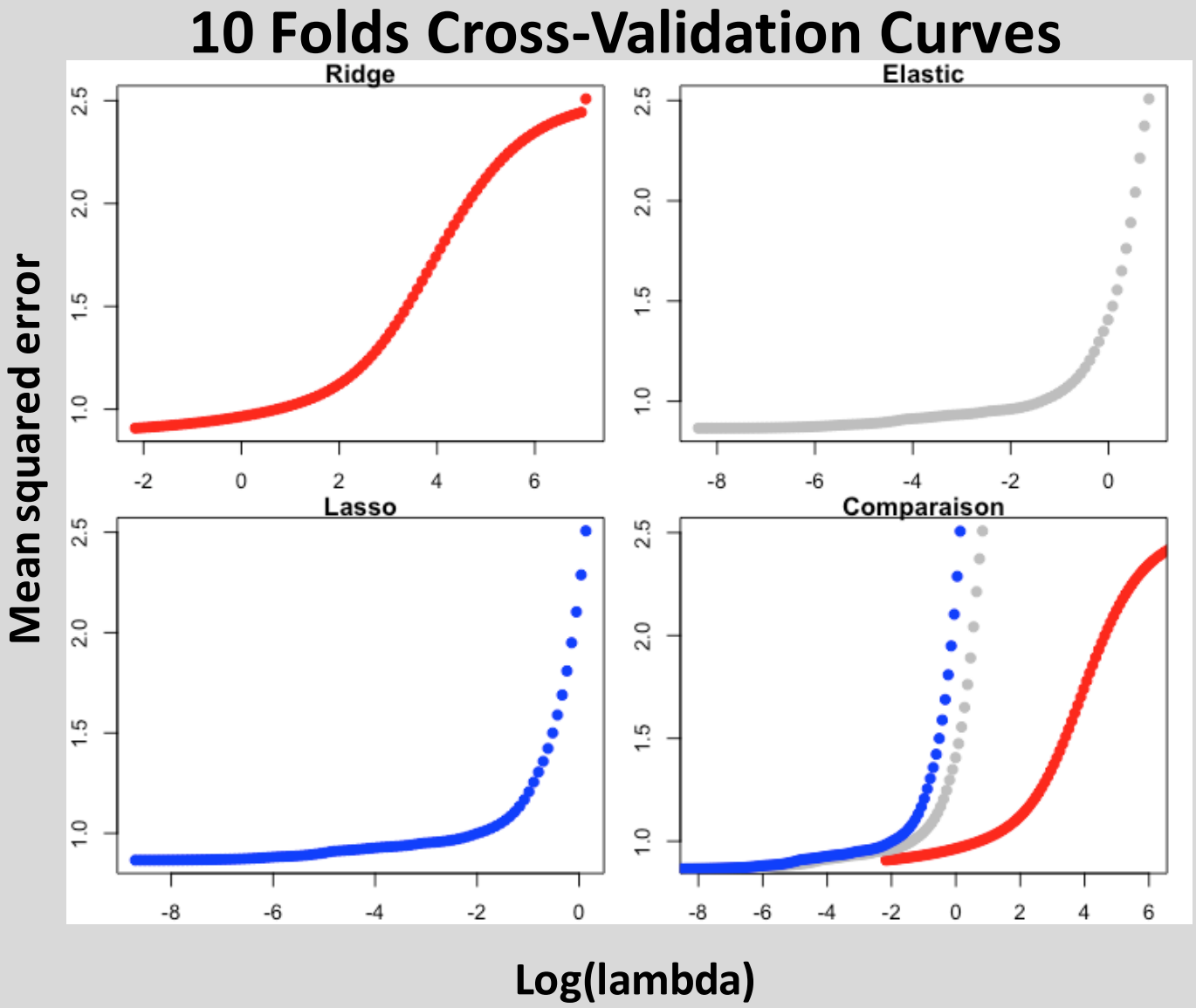
Out of the 81 predictors, 40 were removed as they were the weighted statistic of other predictors

- **Response Variable:**

The original distribution of the critical temperatures had an inverse distribution and thus I took the logarithm for more accurate results

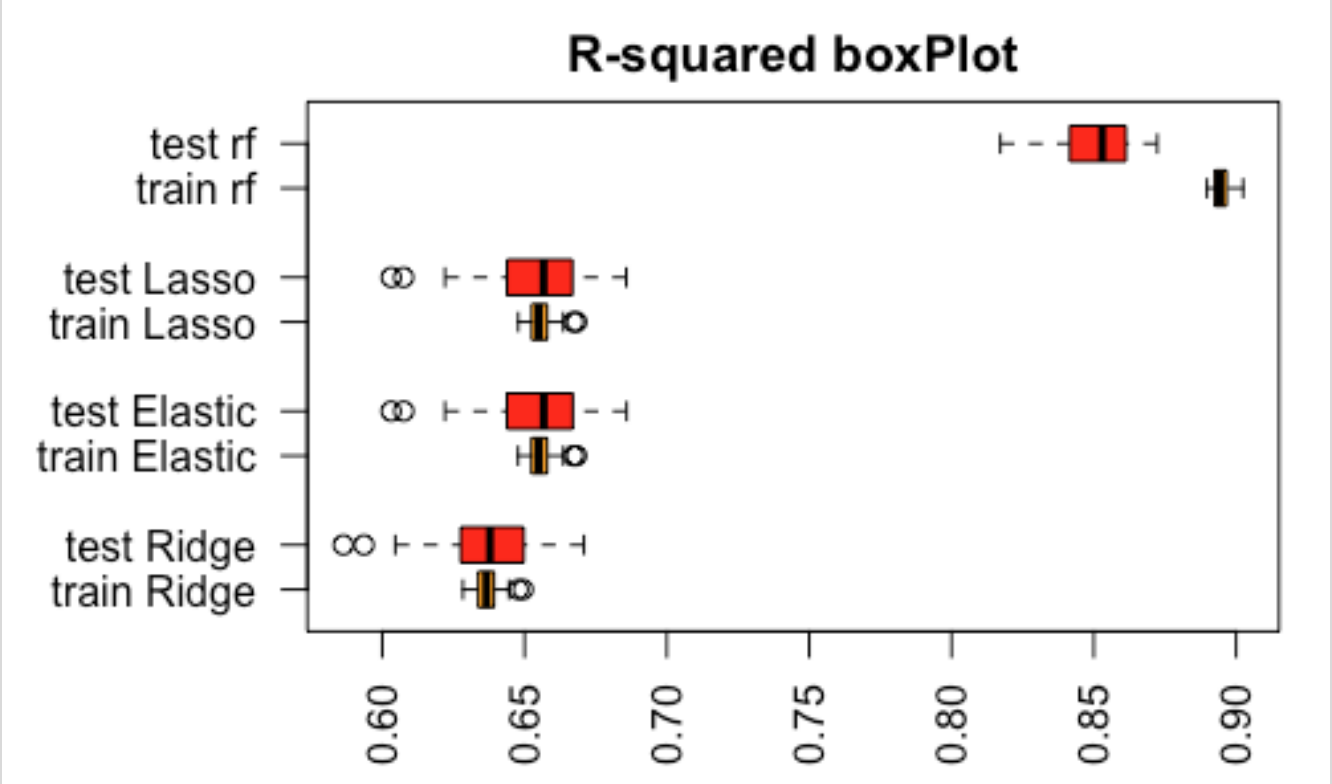
Tuning our lambda for regression

None of the cross-validation curves has a local minimum



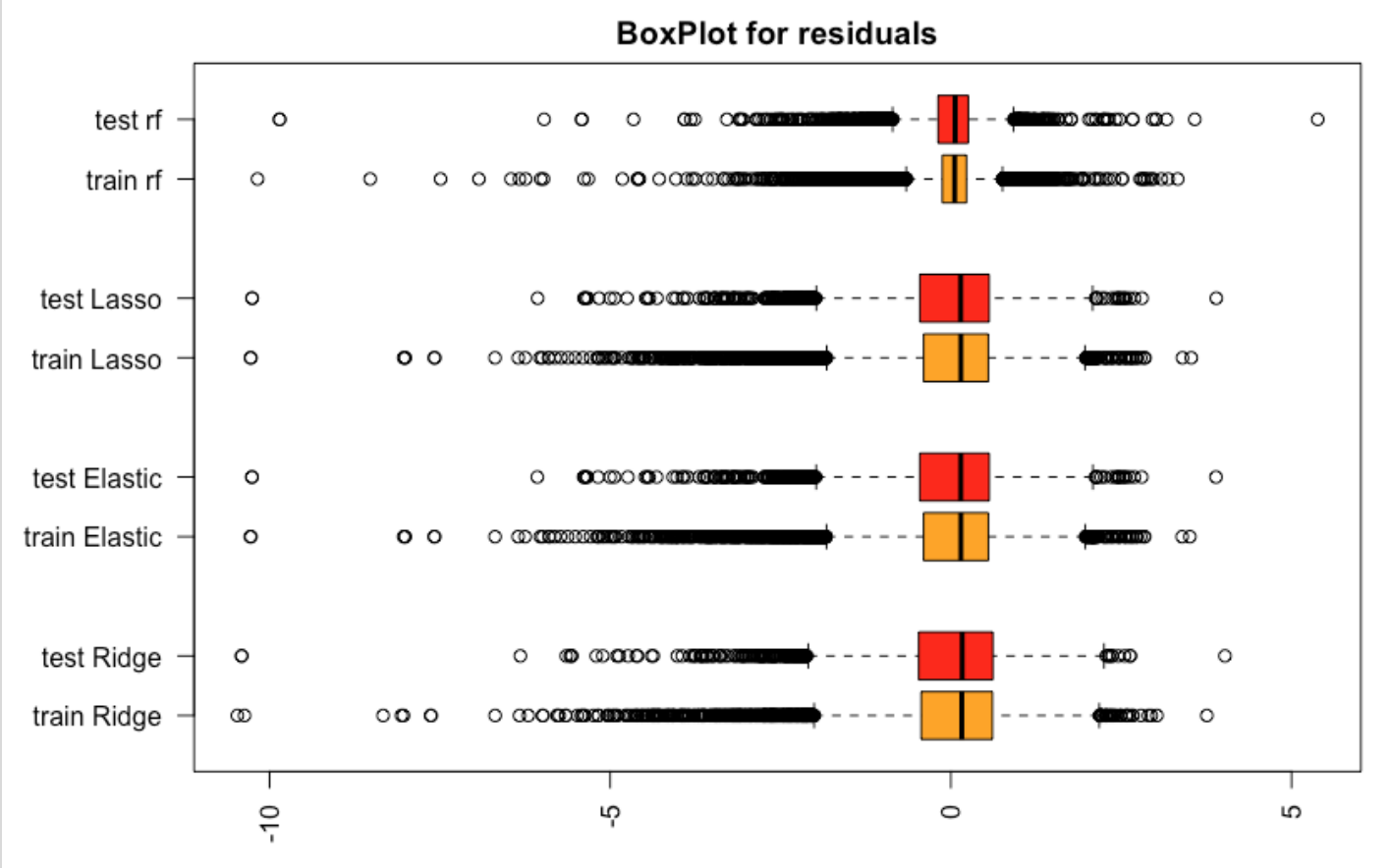
Model Comparison based on R-squared

- **Random Forest seems to be the best model**
- **Always more variation for the test than training data**
- **For regression, test average R-squared is closer to the training one than for random forest**
- **All least-square methods don't seem to do well**



Looking at the residuals

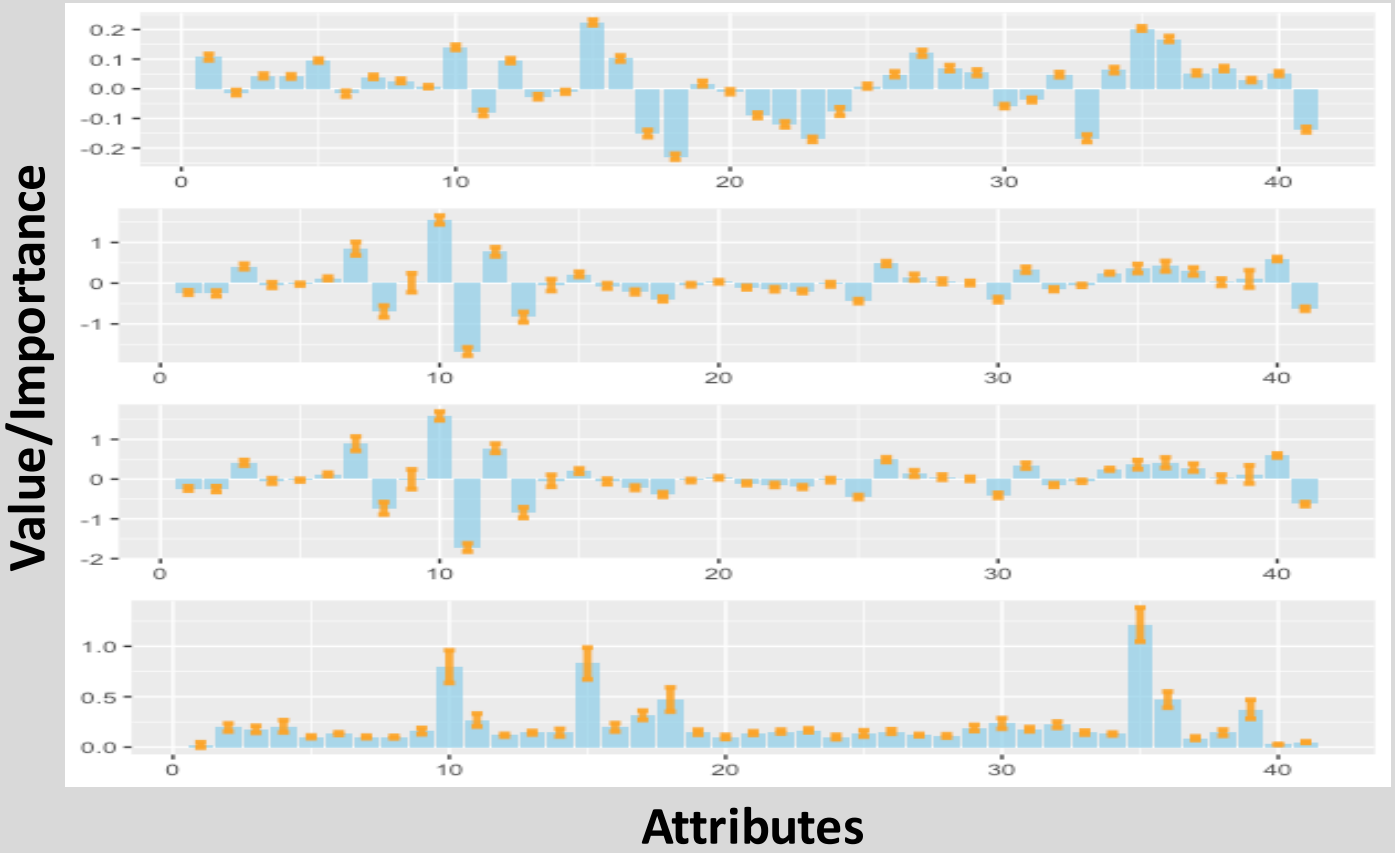
- **The range of residuals for the random forest is smaller than the others**
- **There doesn't seem to be a large difference between test and training**
- **The residuals are well centered and do not seem unbalanced**



Estimation of coefficients

- **Ridge regression has the smallest coefficient**
- **2 of the 3 variables considered important in the rf have small weight in Lasso and Elastic**
- **Range fie, Atomic radius and Thermal conductivity are the most important variables according to the random forest model**

Importance and estimation of coefficients



Summary of the analysis process

- **Random Forest model is the most accurate model**
- **Regression overall doesn't seem to be efficient to predict the critical temperature of superconductors**
- **Ridge, Elastic-net and Lasso have close to the same run time and a log complexity**
- **Random Forest seem to have an exponential complexity**
- **Overall classification seems to be a better method to estimate critical temperature**

	Ridge	Elastic-net	Lasso	Random Forest
N=2000	0.25 sec	0.43 sec	0.41 sec	0.66 sec
N=5000	0.39 sec	0.63 sec	0.54 sec	2.39 sec
N=21263	0.98 sec	0.95 sec	0.97 sec	20.25 sec
MSE	0.90	0.86	0.86	0.38