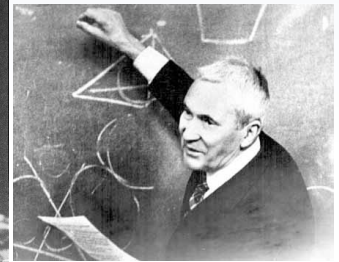
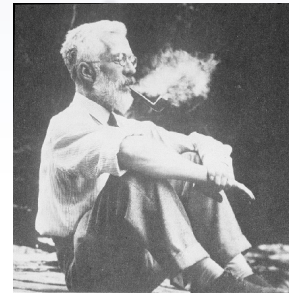
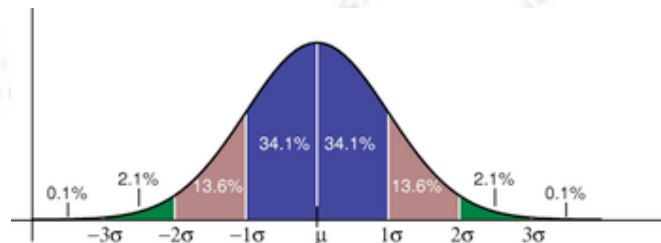


# Big Data Analysis

## Results and Scores of Small Project



Troels C. Petersen & Carl Johnsen (NBI)



*“Statistics is merely a quantisation of common sense - Big Data is a sharpening of it!”*

# Overall comments

The name “Small Project” is misleading, and should have been “Initial project”, because it is by no means small. But you did very well, and so let me start by gently stating, that you have little / nothing to fear - in fact, you did really great!

Grading it was perhaps harder than the project itself, but Carl and I have done our best to be as open as possible about the scoring. And to give you a maximum of feedback, we have produced a report for each of you.

# The motivation

We wanted you to try the very **real challenge** of optimising models, without knowing their performance on the data it is applied to.

We also wanted you to **individually** run ML algorithms, so that you have the machinery in place after the course.

We insisted that you tried **both tree- and NN-based algorithms**, to get a feel for their differences and similarities.

The description file was meant to trigger you to **think about your models**, and what you tried. Also, considerations of size and performance are in place.

Finally, we wanted to **ensure** that you yourself tried all the work and things to consider, to put together ML models and apply them.

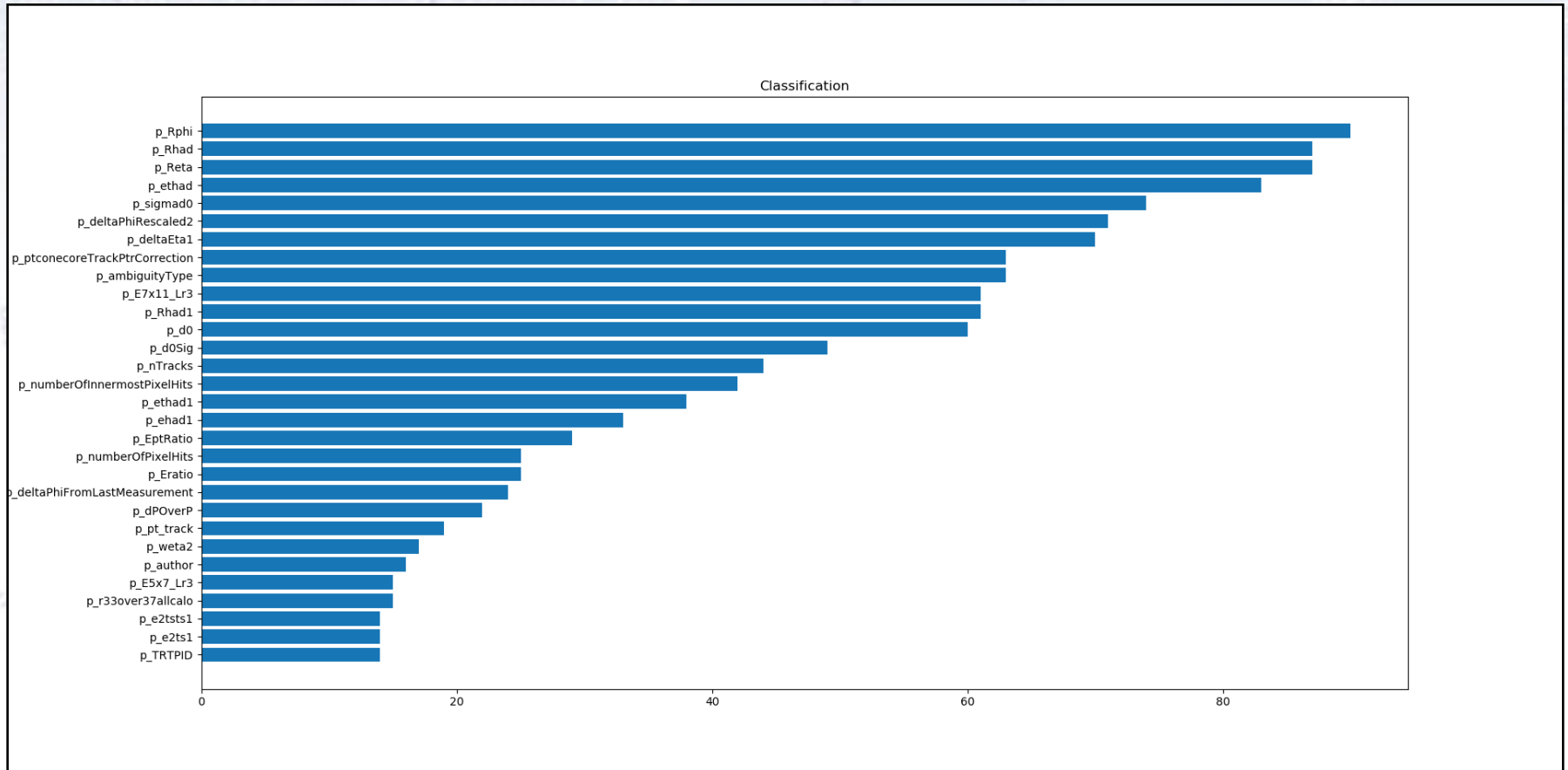


# Classification Results



# Classification variable usage

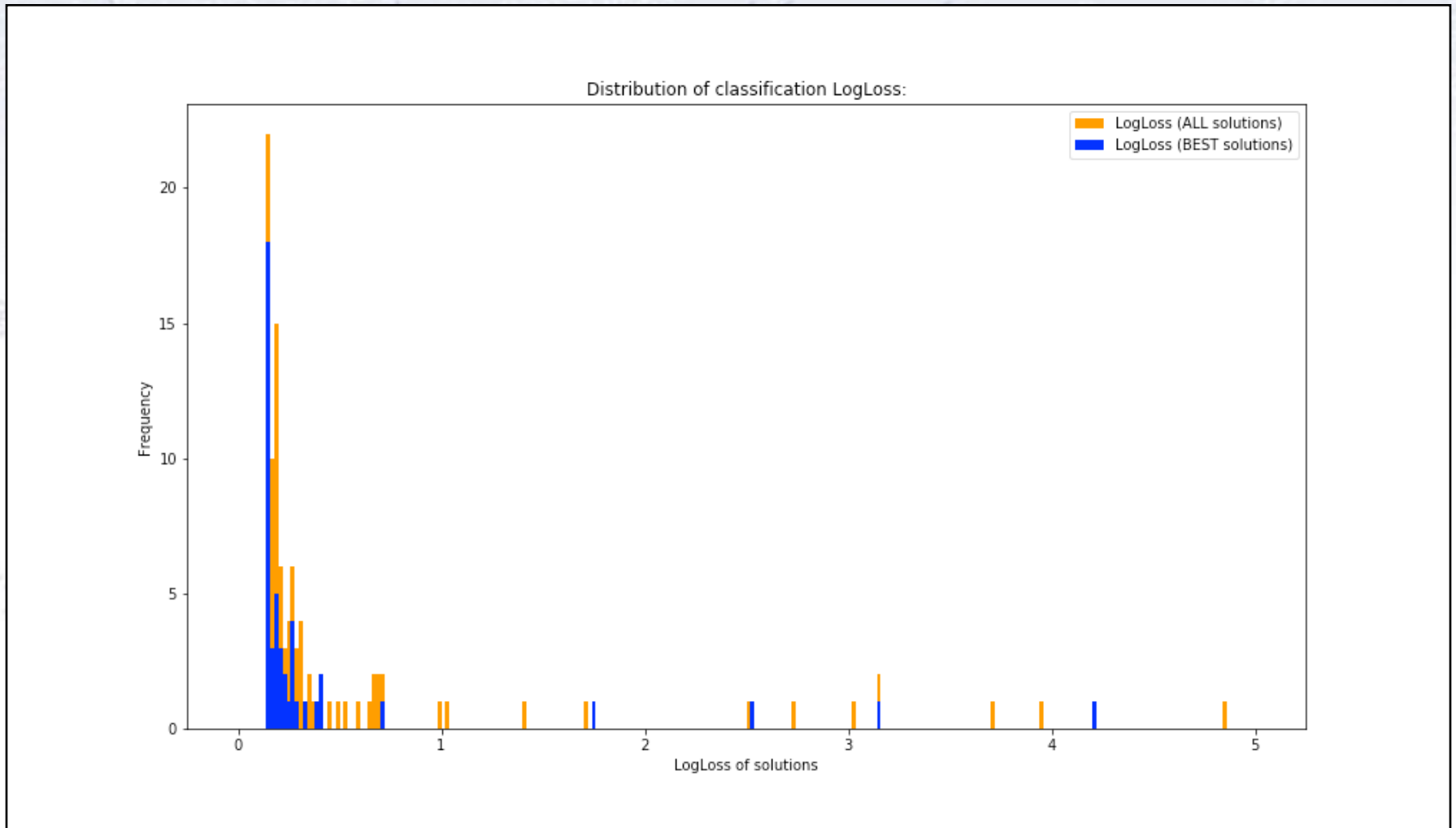
Many (most?) of you have made a variable ranking. Below you find a variable usage frequency plot, showing how often a variable was used.



There is a small “step” after 12 variables, which probably reflects the output of different variable ranking methods (permutation importance and SHAP).

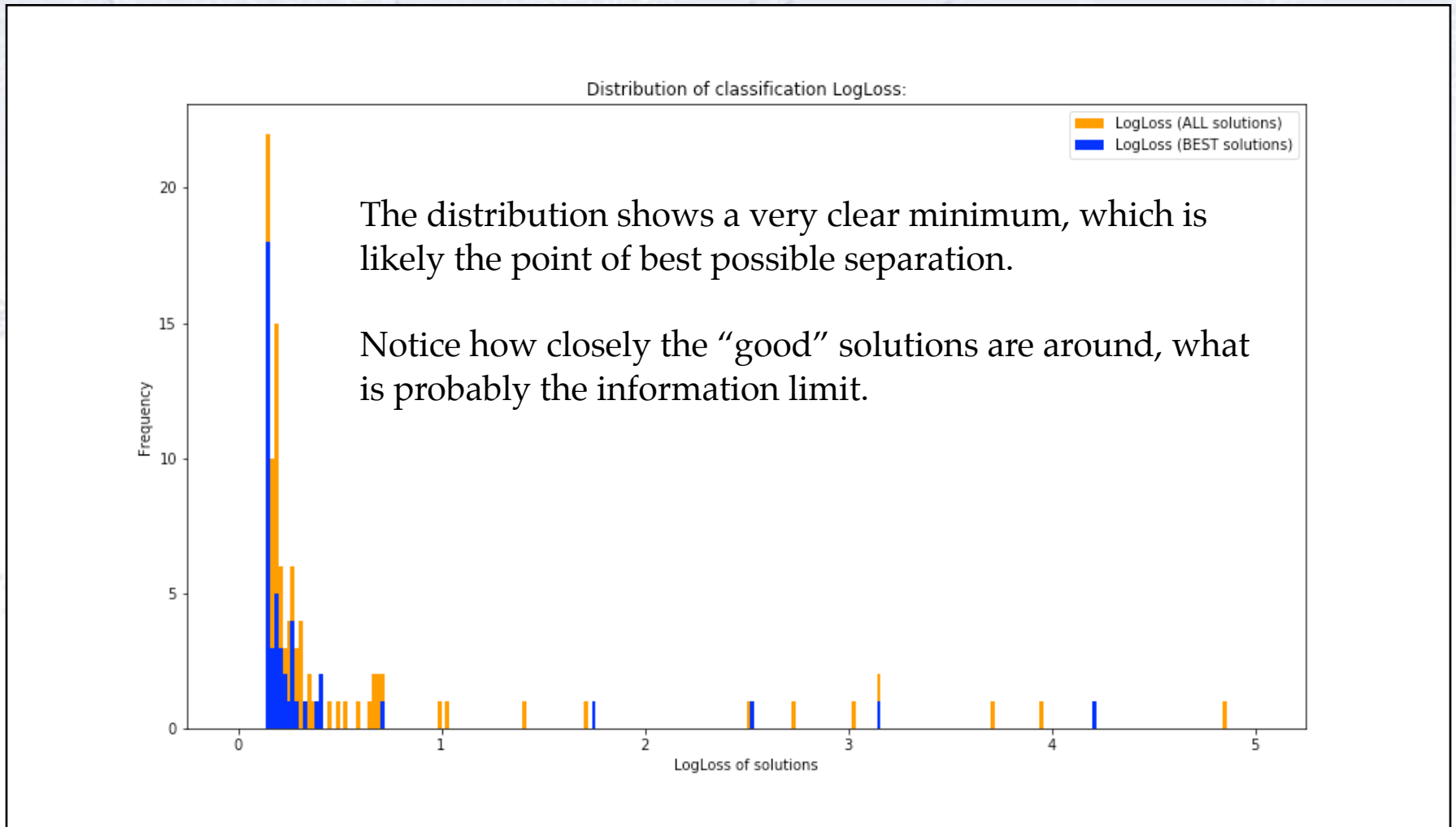
# Classification score distribution

The distribution of the (Cross-Entropy) LogLoss values obtained was:



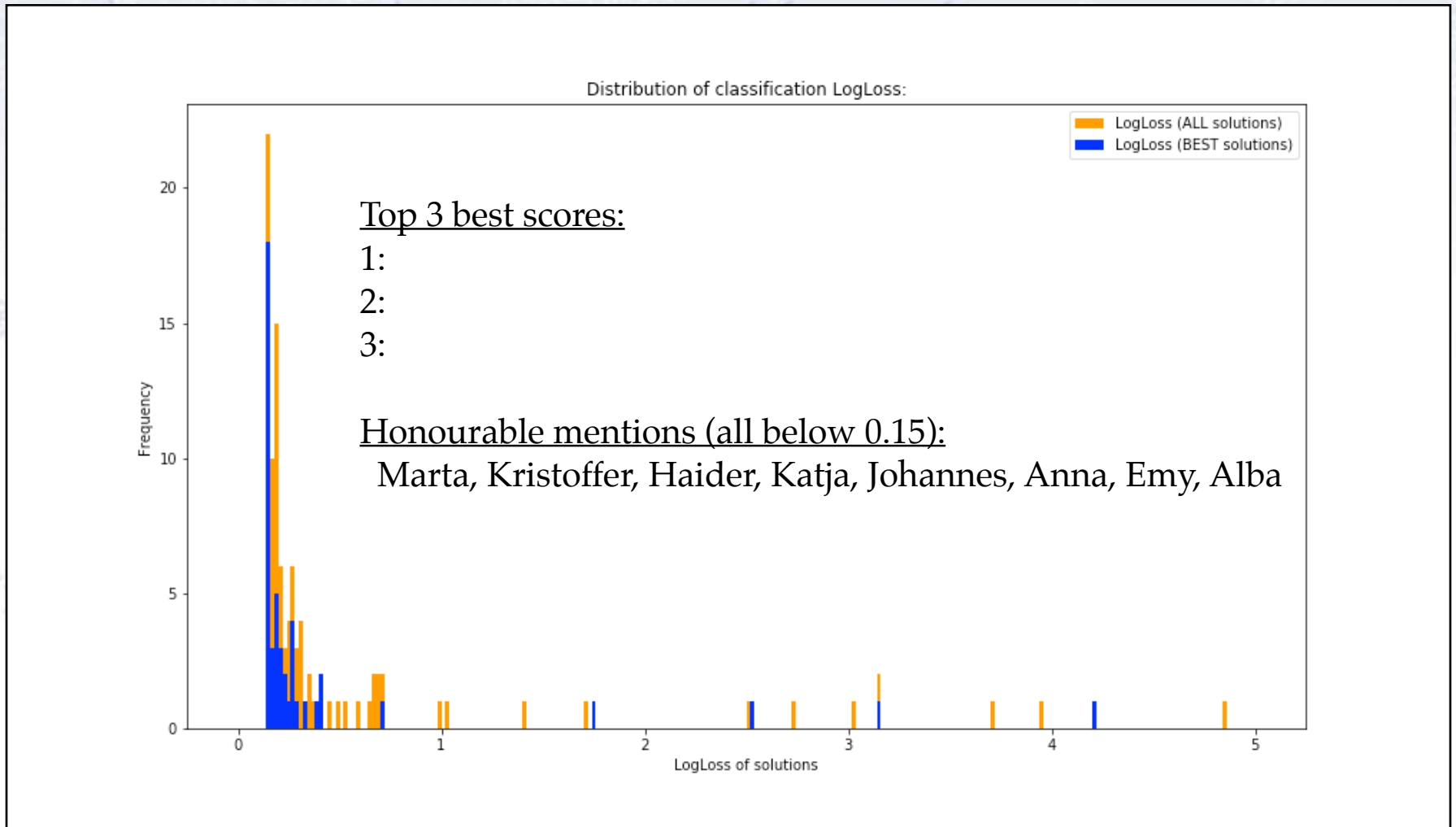
# Classification score distribution

The distribution of the (Cross-Entropy) LogLoss values obtained was:



# Classification score distribution

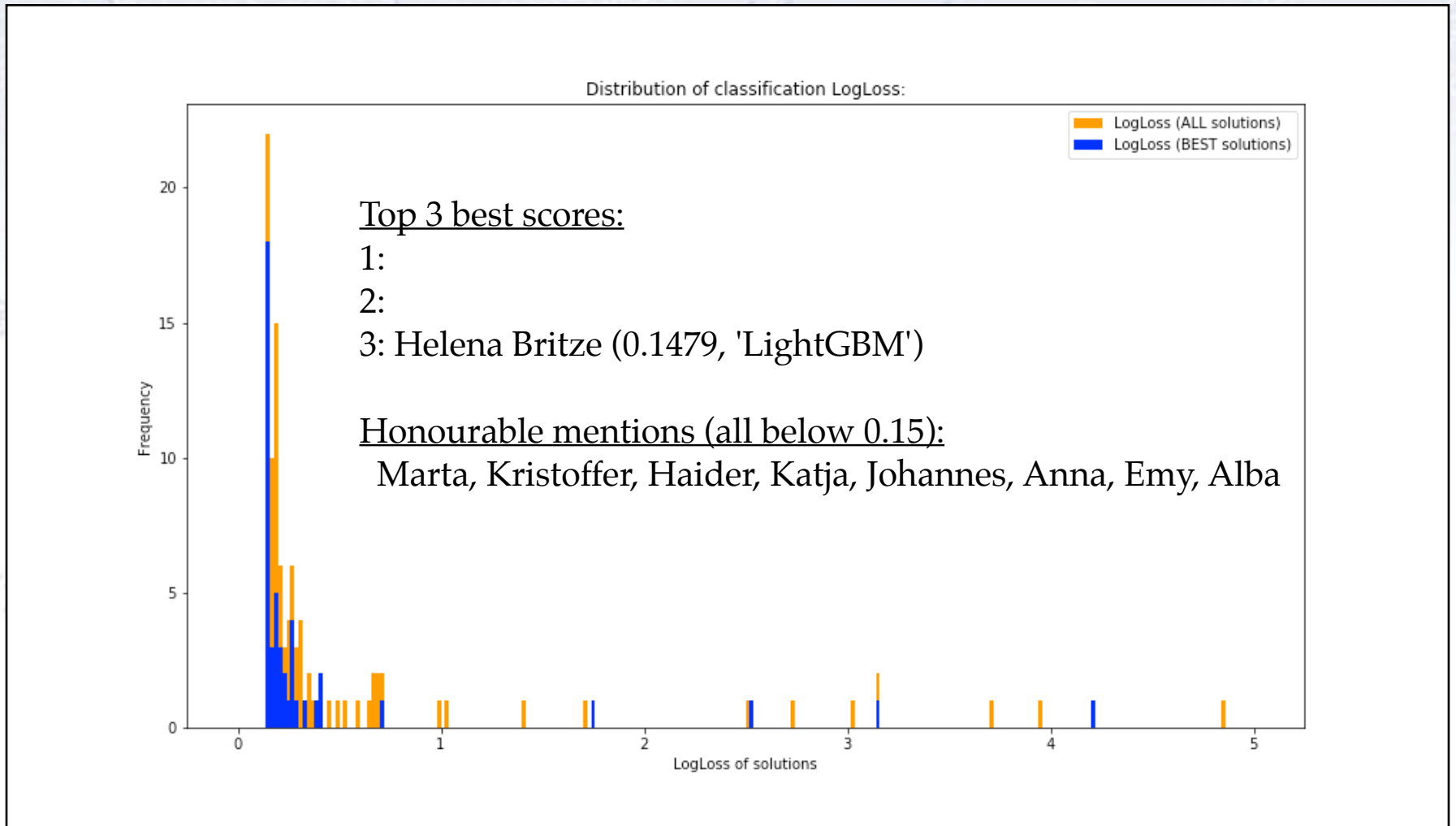
The distribution of the (Cross-Entropy) LogLoss values obtained was:





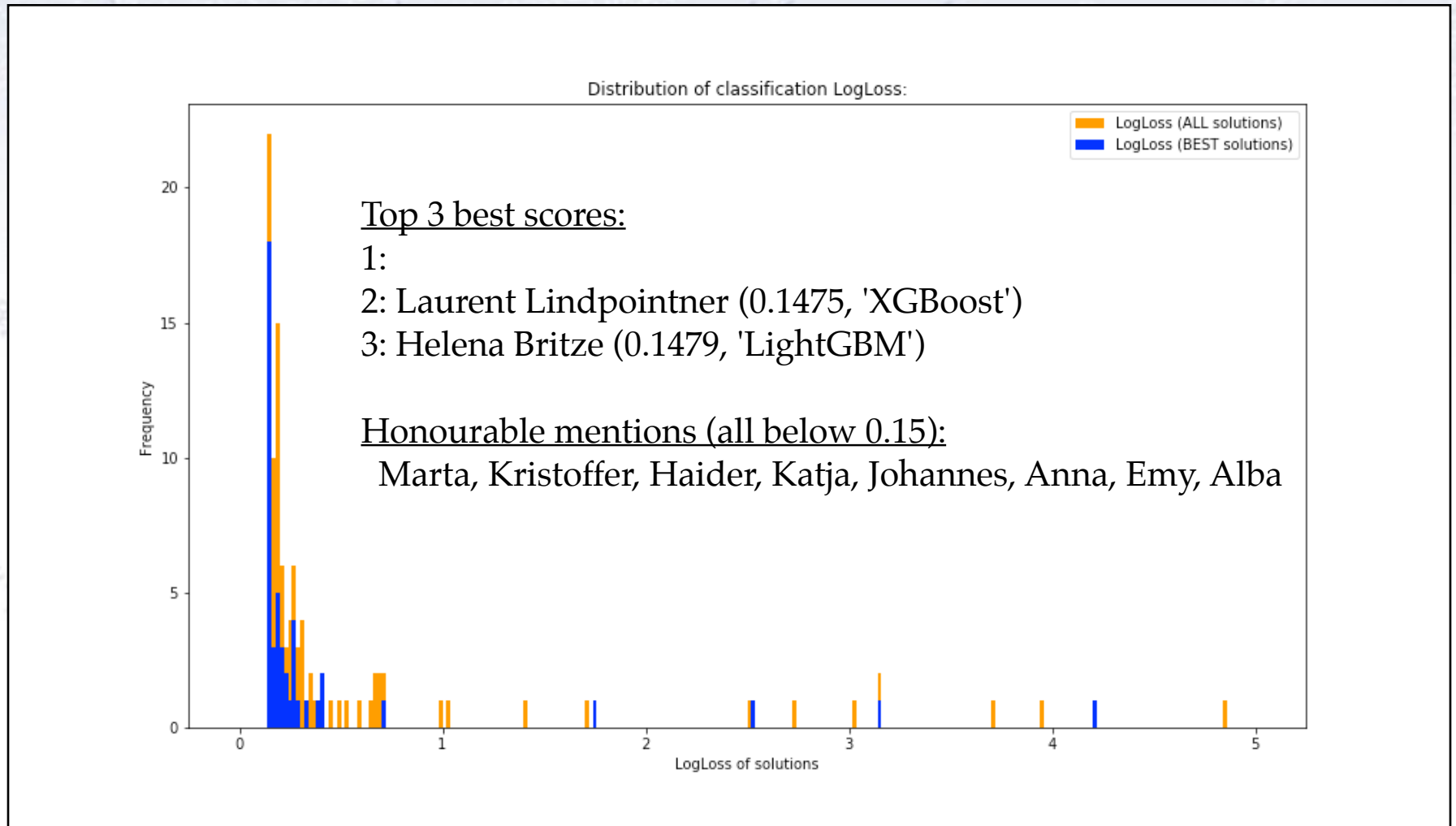
# Classification score distribution

The distribution of the (Cross-Entropy) LogLoss values obtained was:



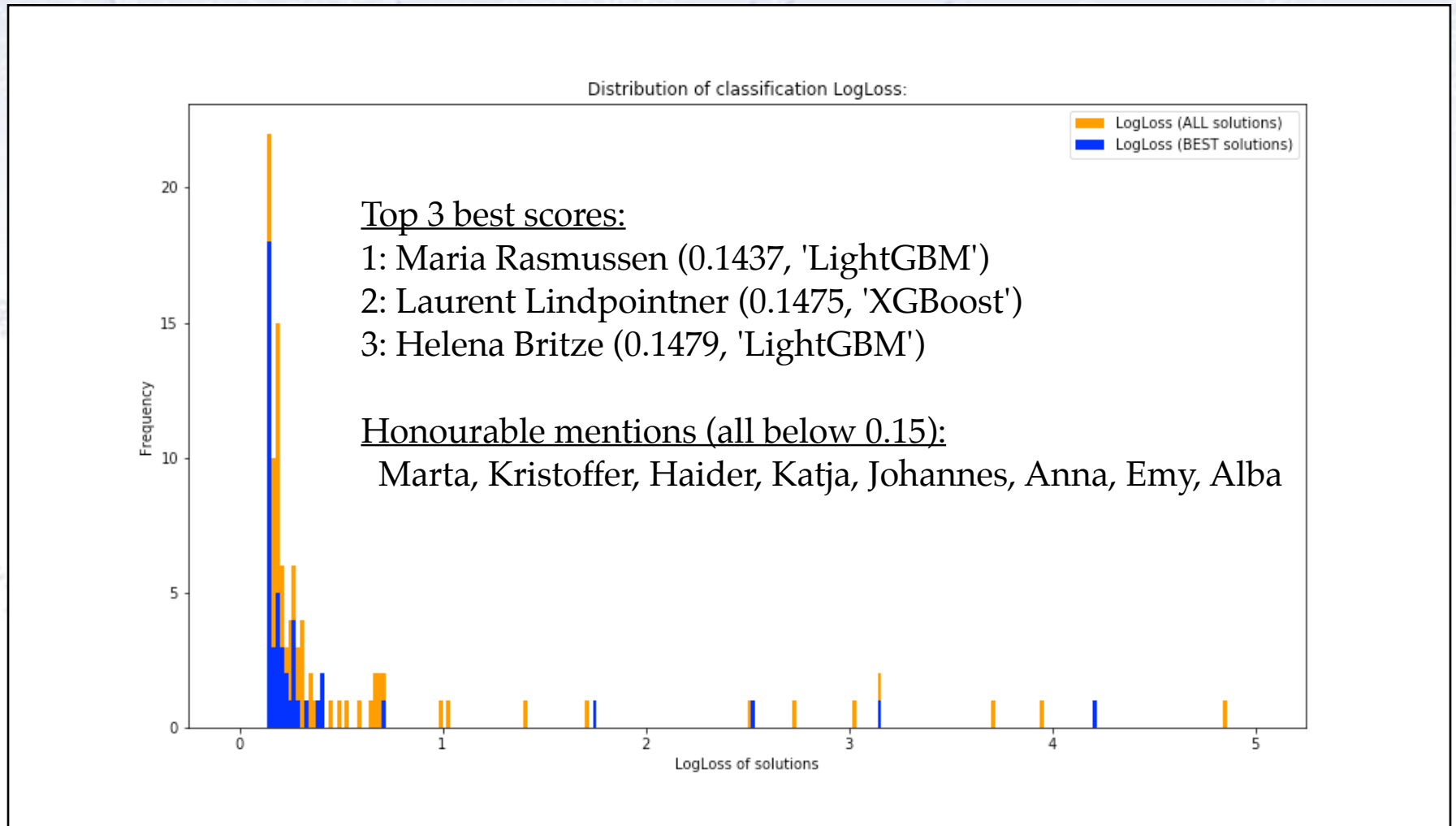
# Classification score distribution

The distribution of the (Cross-Entropy) LogLoss values obtained was:



# Classification score distribution

The distribution of the (Cross-Entropy) LogLoss values obtained was:

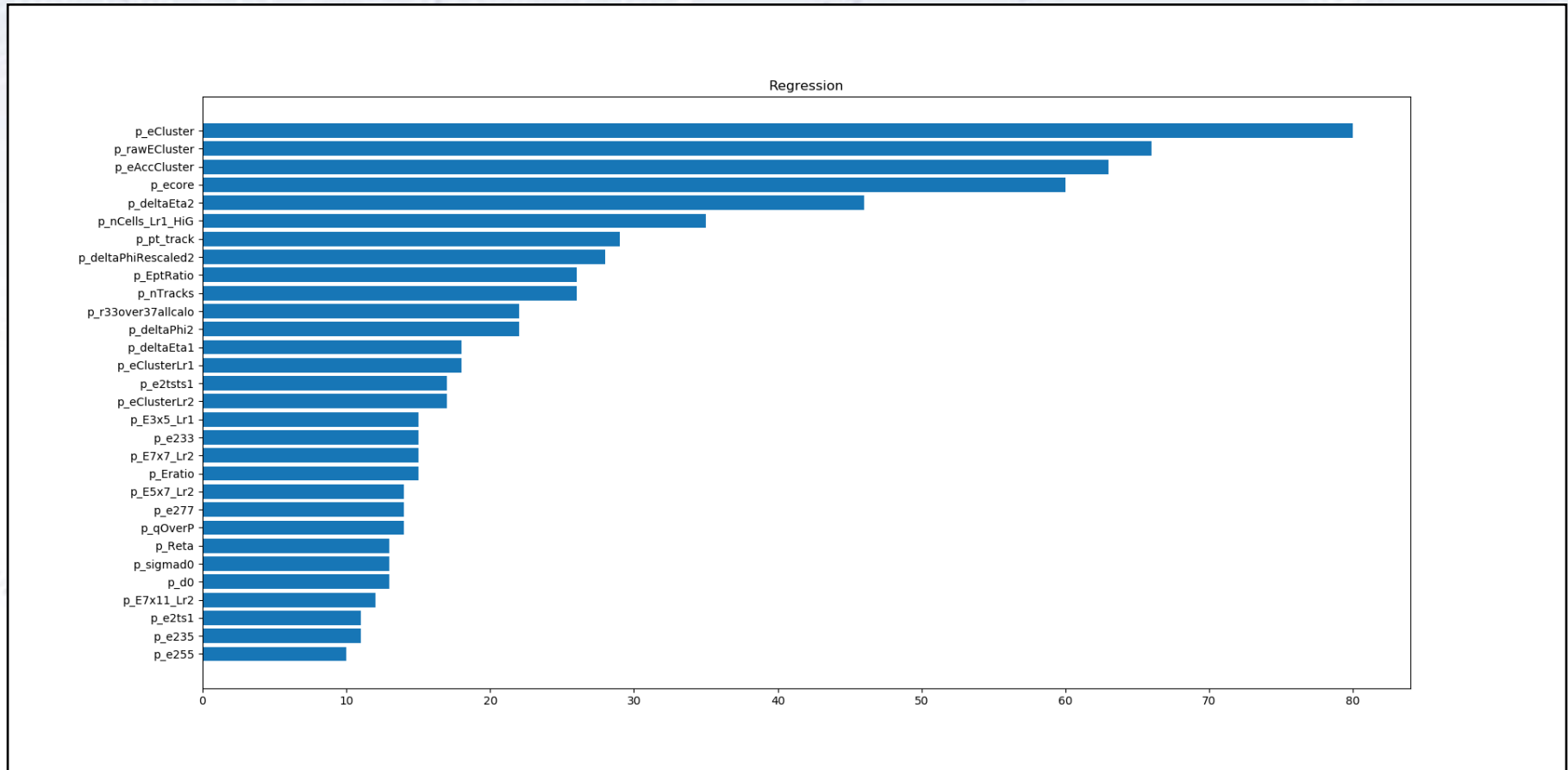


A faded background image of a nautical chart. The chart features concentric circles representing magnetic isotherms, with values ranging from 0 to 360 degrees. A compass rose is visible in the center, and the word "MAGNETIC" is printed on the chart. The chart is overlaid with a grid of latitude and longitude lines. The text "REGRESSION RESULTS" is prominently displayed in the center of the chart.

# Regression Results

# Regression variable usage

The most important variable happens to be ATLAS' own energy prediction, so that is no surprise. I considered not including it, and might change next year.

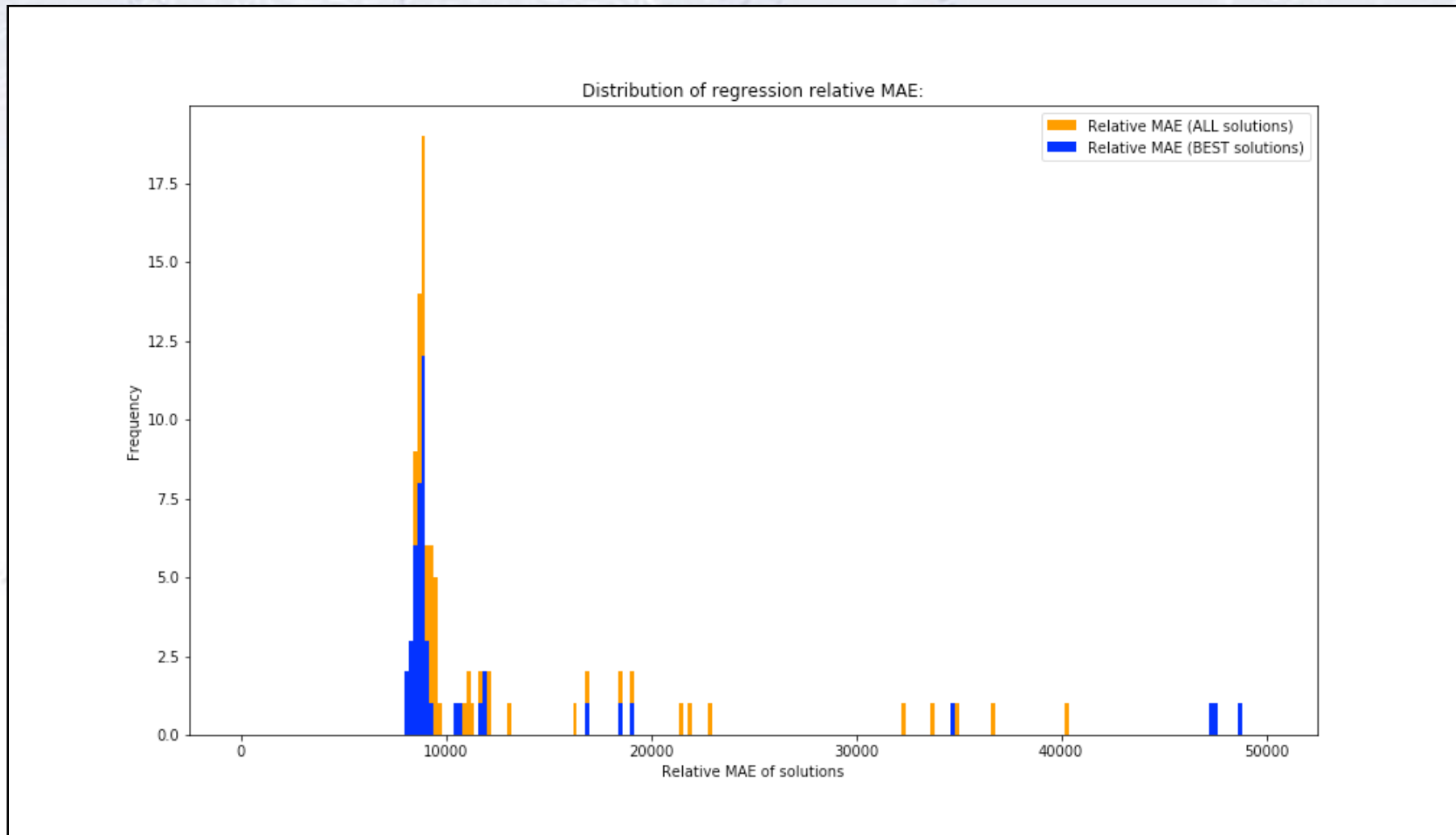


The variables have changed drastically from the PID case, and there is NO overlap at all for the top 10-15 variables! PID and E-regression are two very different tasks.



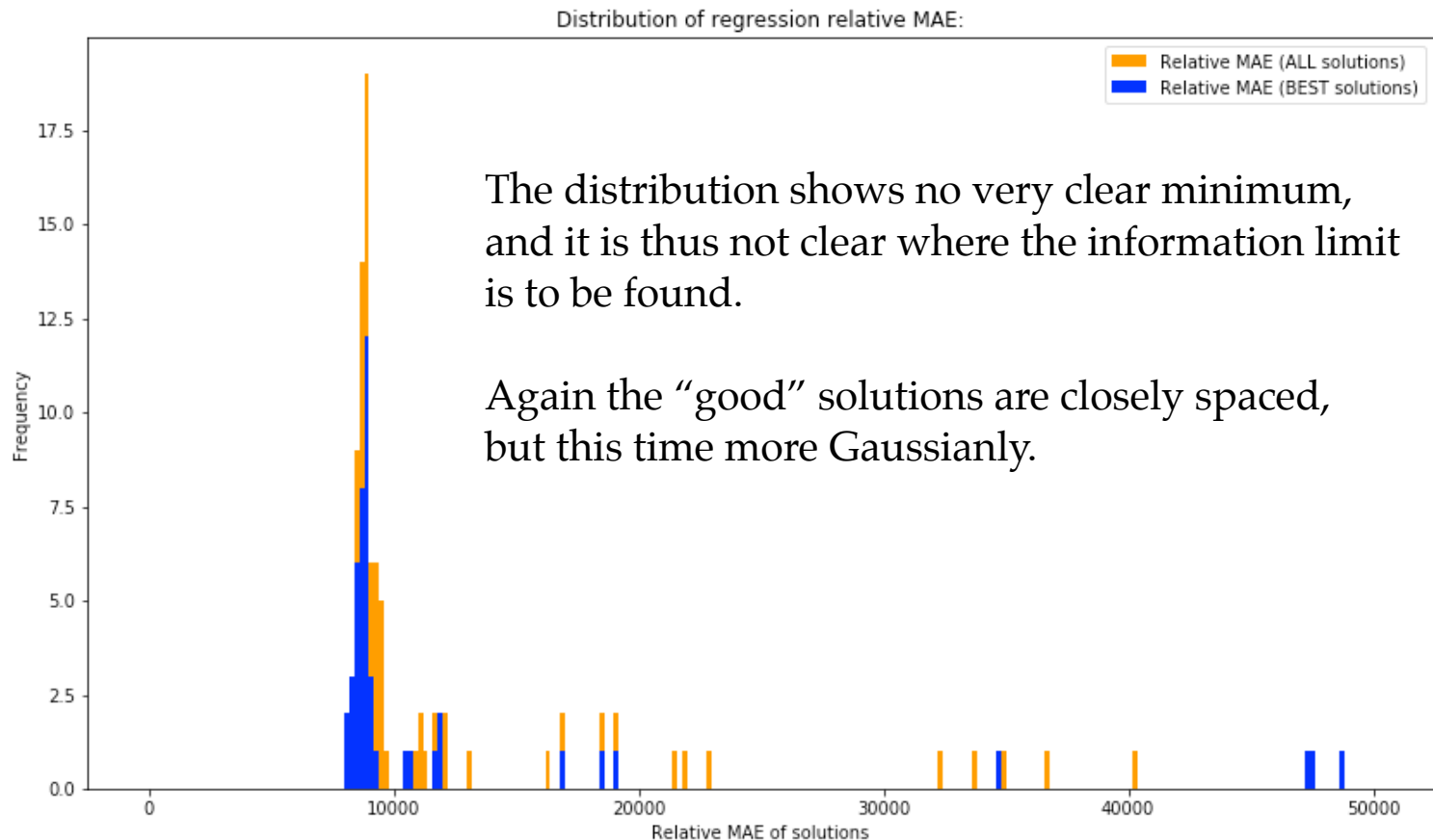
# Regression score distribution

The distribution of the relative MAE (i.e.  $MAE((E-T)/T)$ ) values obtained was:



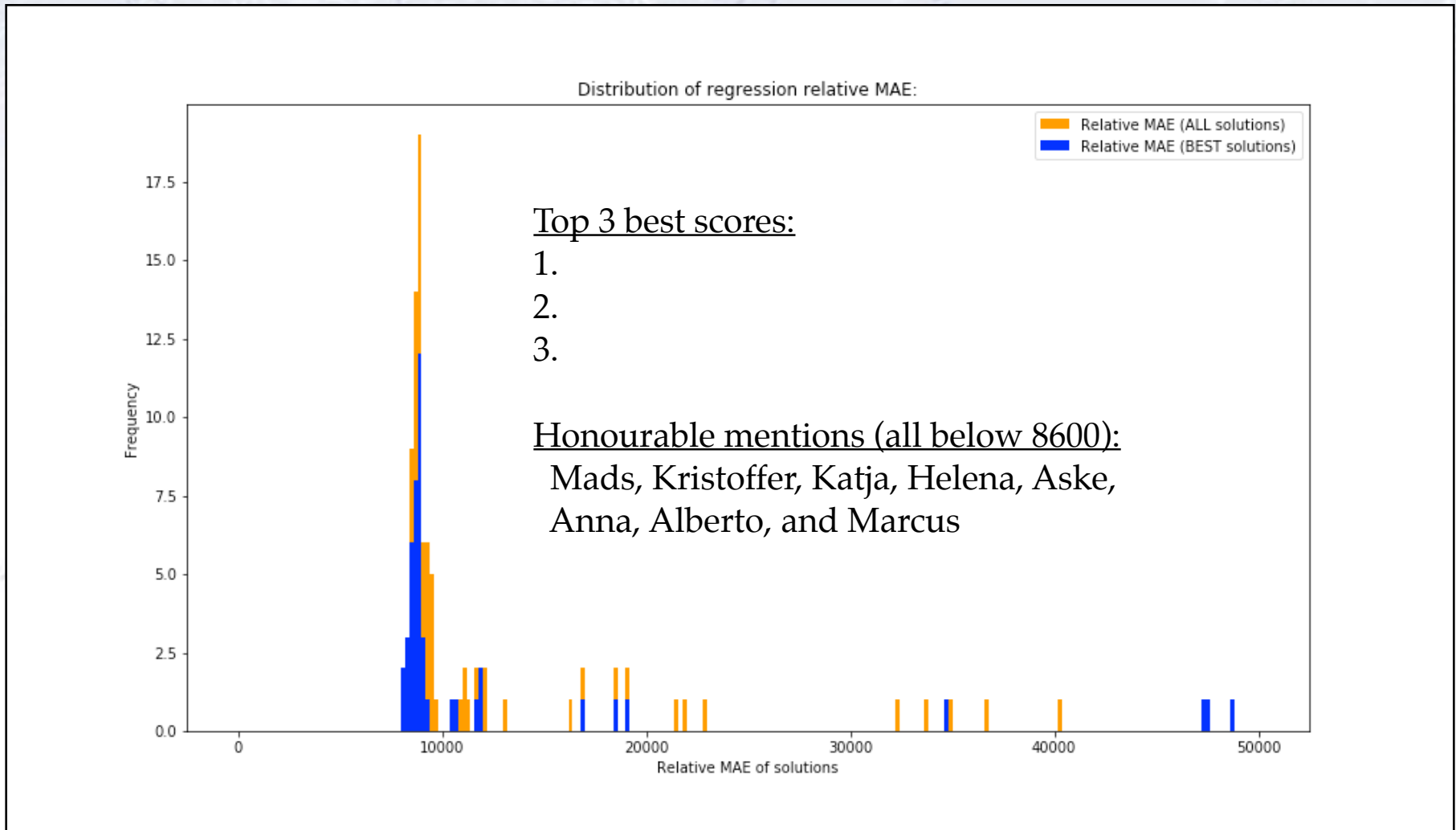
# Regression score distribution

The distribution of the relative MAE (i.e.  $\text{MAE}((E-T)/T)$ ) values obtained was:



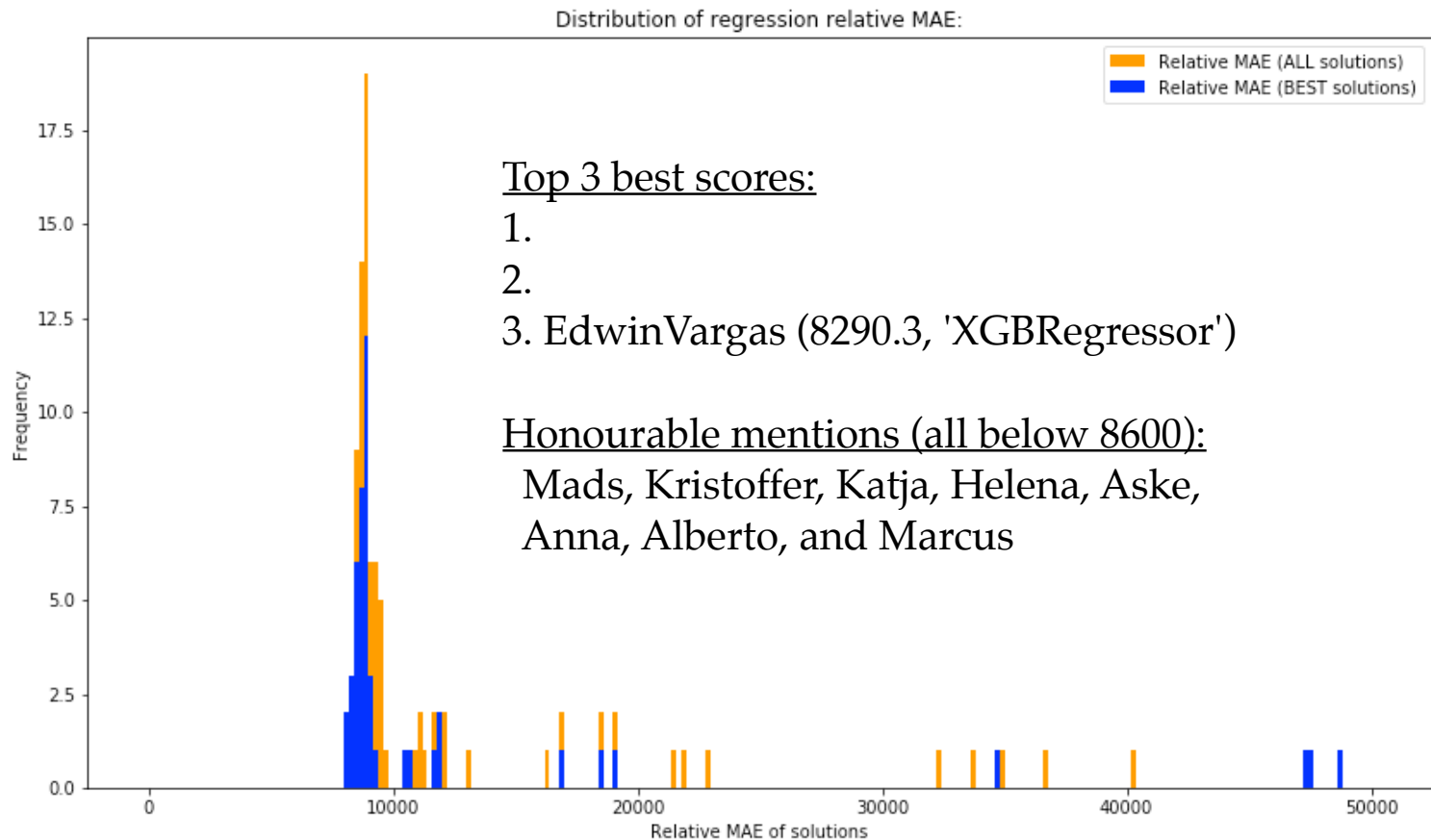
# Regression score distribution

The distribution of the relative MAE (i.e.  $MAE((E-T)/T)$ ) values obtained was:



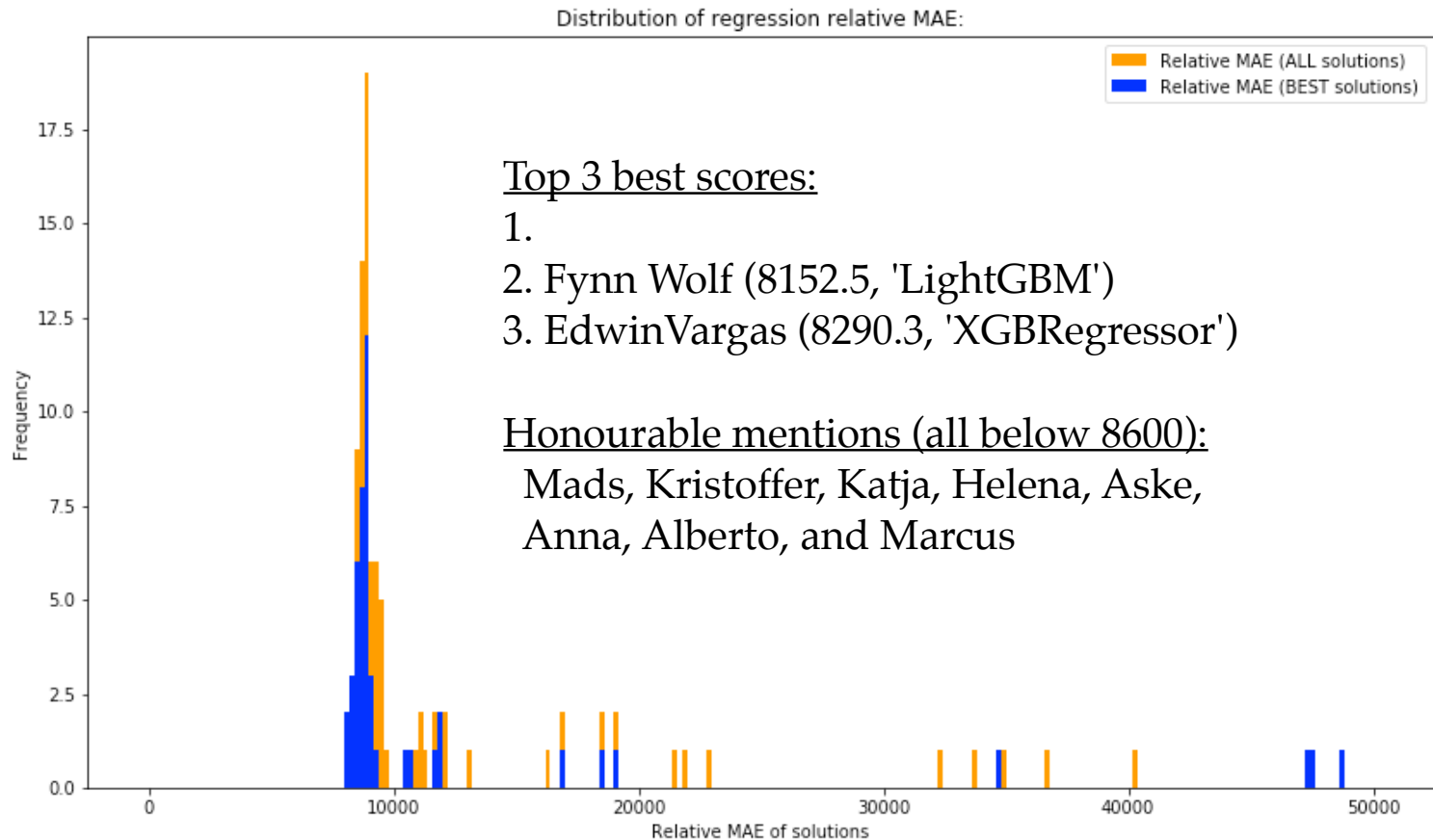
# Regression score distribution

The distribution of the relative MAE (i.e.  $MAE((E-T)/T)$ ) values obtained was:



# Regression score distribution

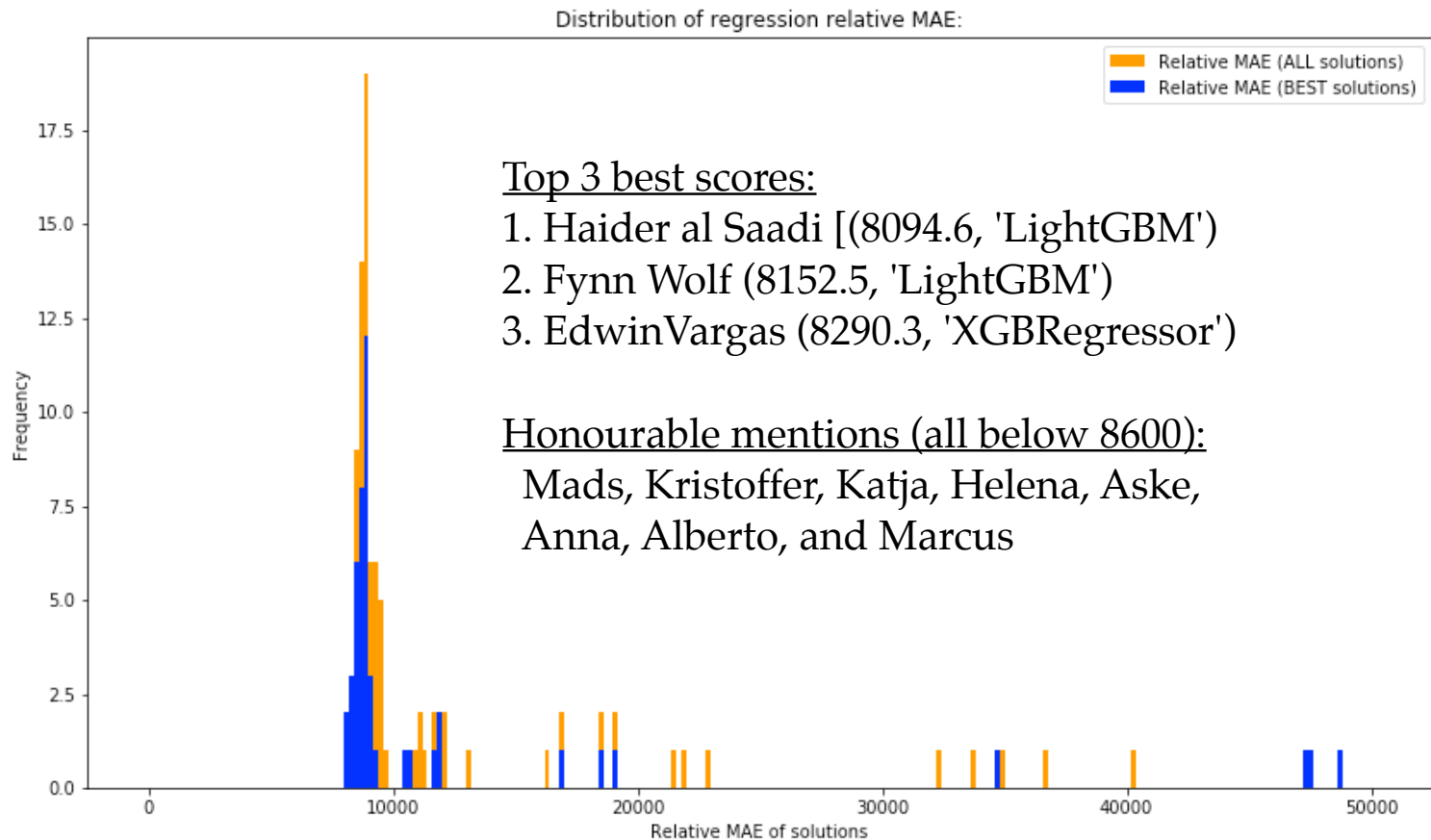
The distribution of the relative MAE (i.e.  $MAE((E-T)/T)$ ) values obtained was:





# Regression score distribution

The distribution of the relative MAE (i.e.  $MAE((E-T)/T)$ ) values obtained was:

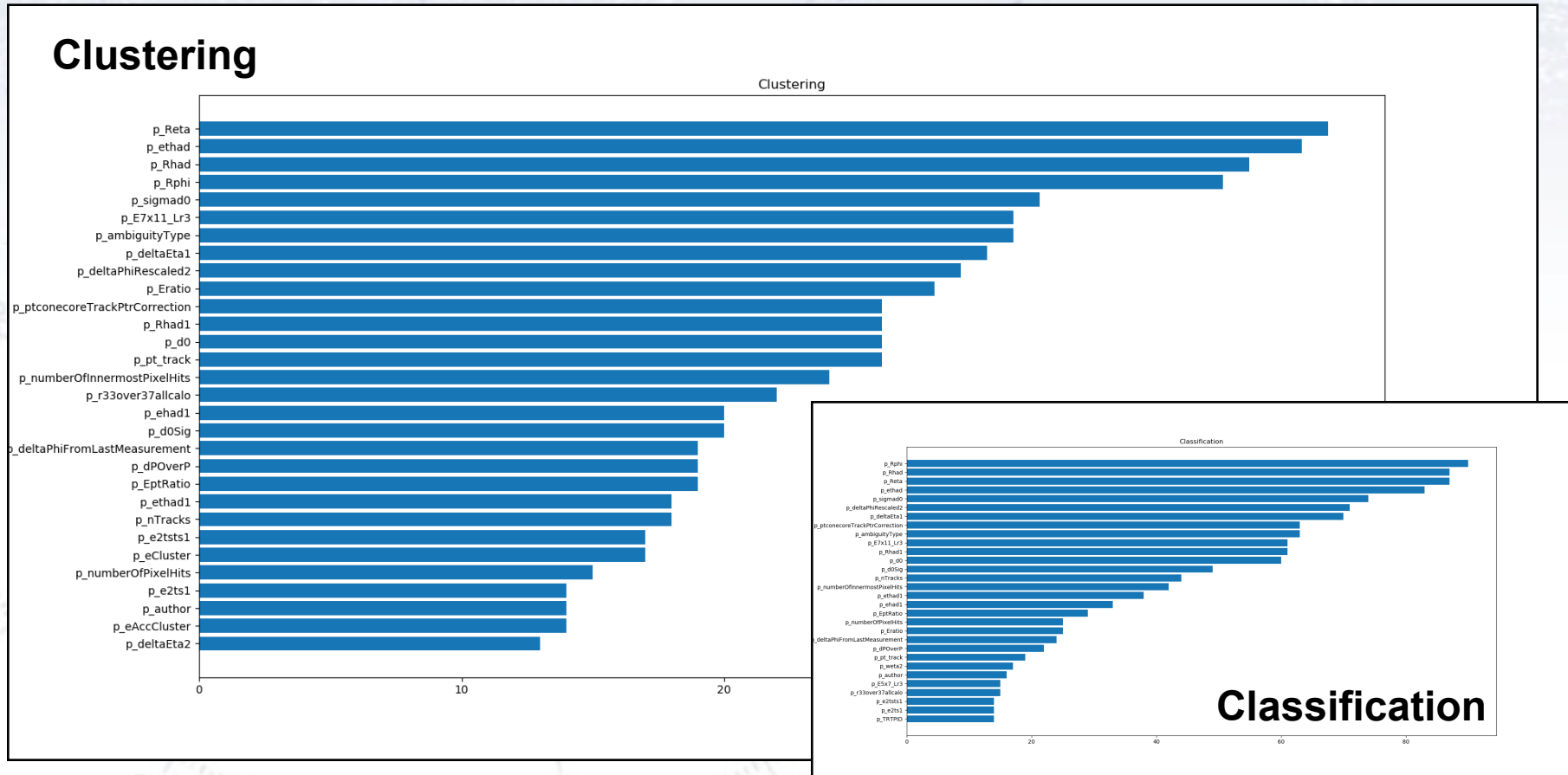




# Clustering Results

# Clustering variable usage

I would have thought, that the clustering variable usage would be near-identical to that of the (supervised) classification task. However, it is not entirely...



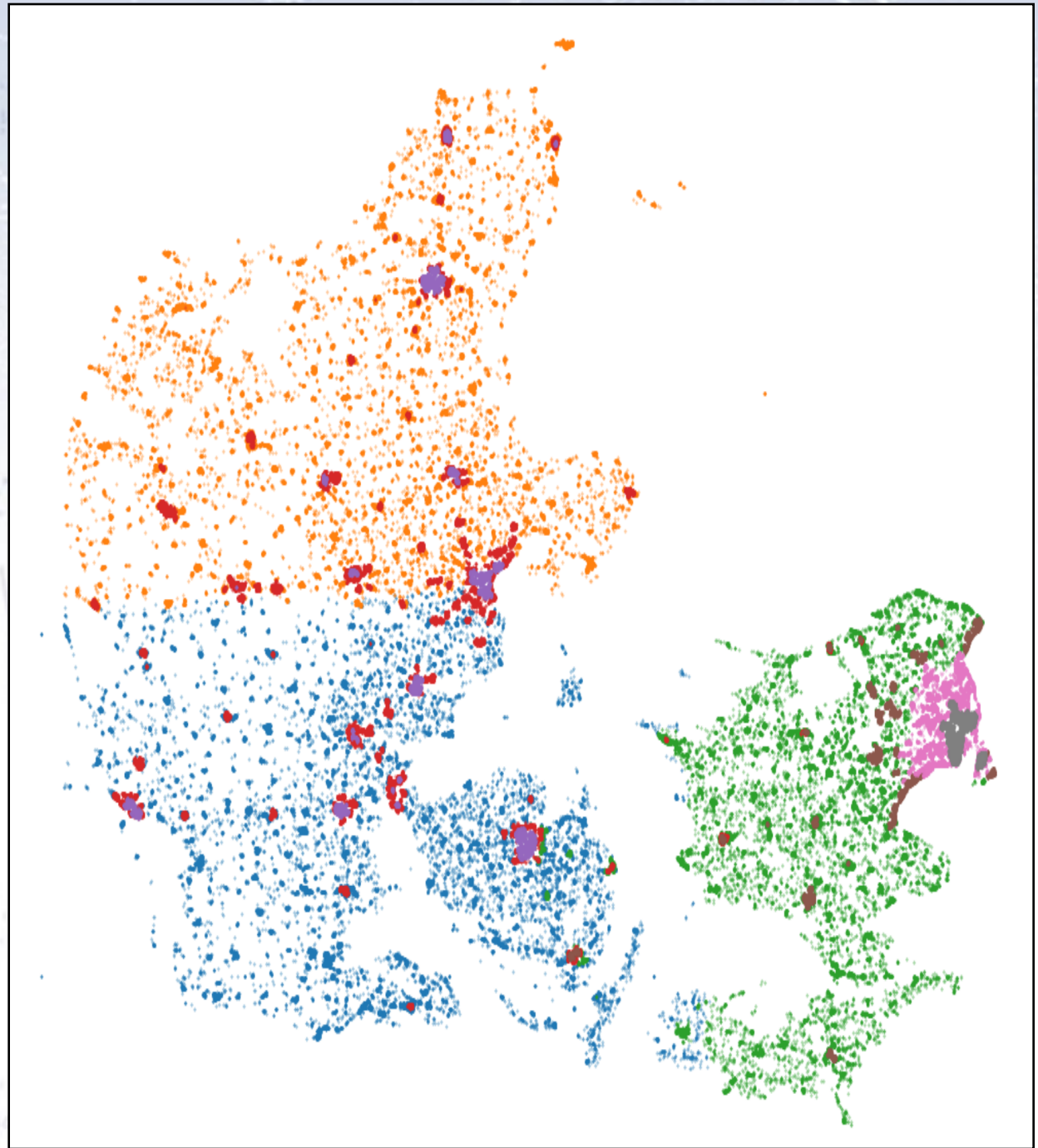
It is also a “hard” (i.e. under defined) task of choosing variables for clustering, when the task / target is unknown. It takes insight and domain knowledge...

# Clustering housing

While postal codes are good, they are not very useful in clustering Denmark.

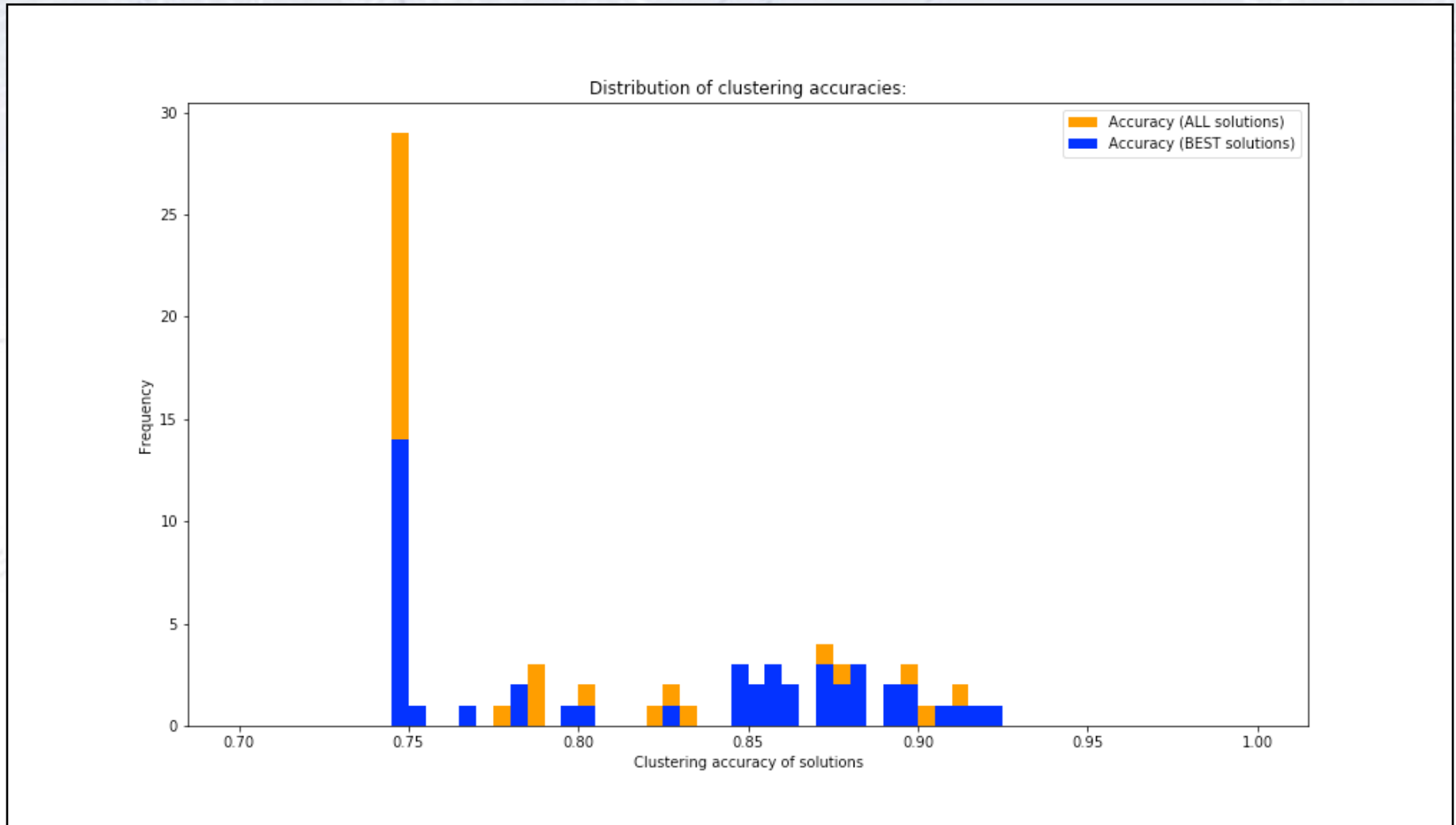
However, using just a few variables ( $x$ ,  $y$ , density, price/m<sup>2</sup>), one can cluster villas in Denmark very efficiently.

In this way, one can follow trends for a type of house much better.



# Clustering accuracy distribution

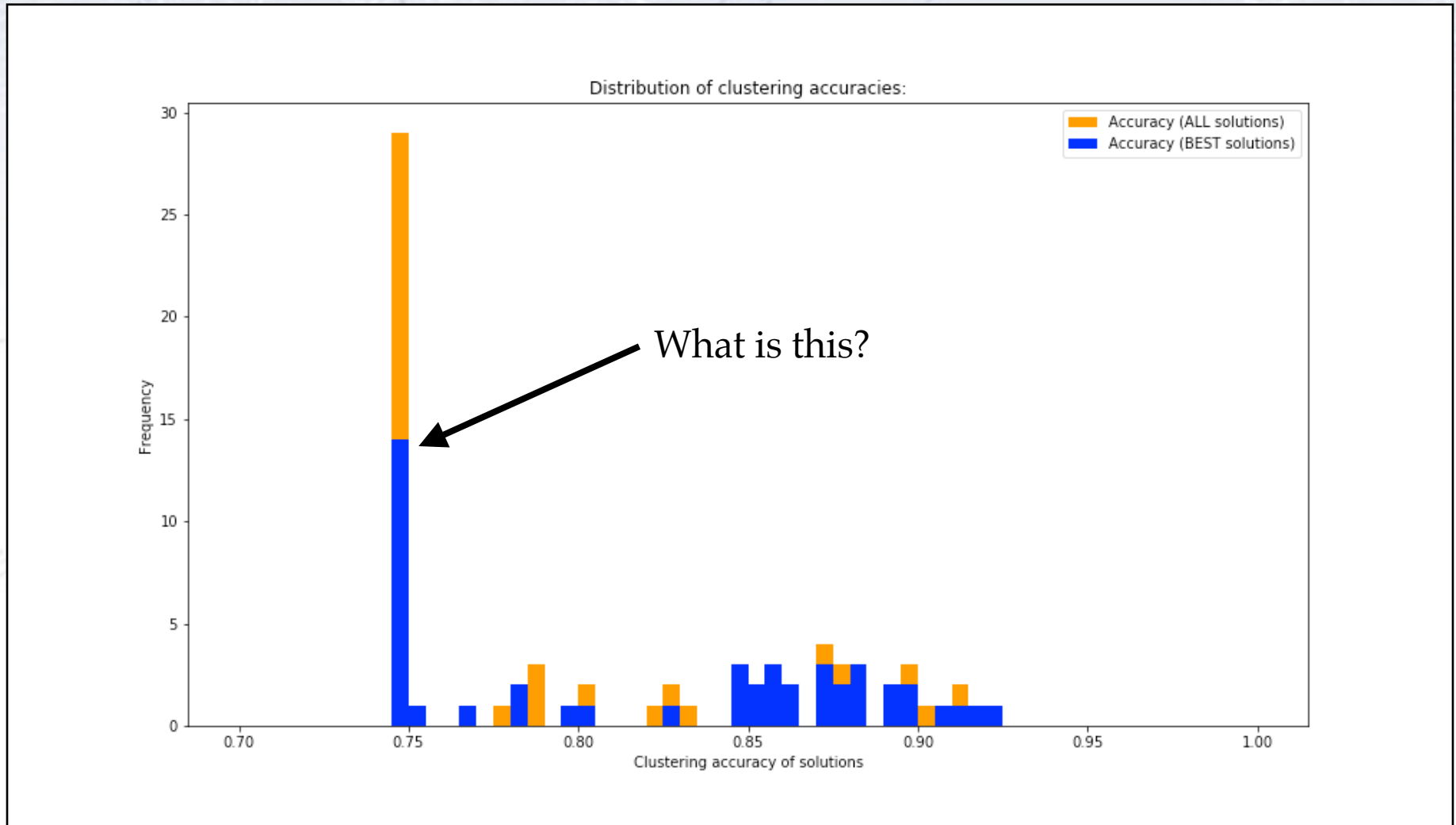
The accuracy of the clustering (when assigned either electron or not) was:





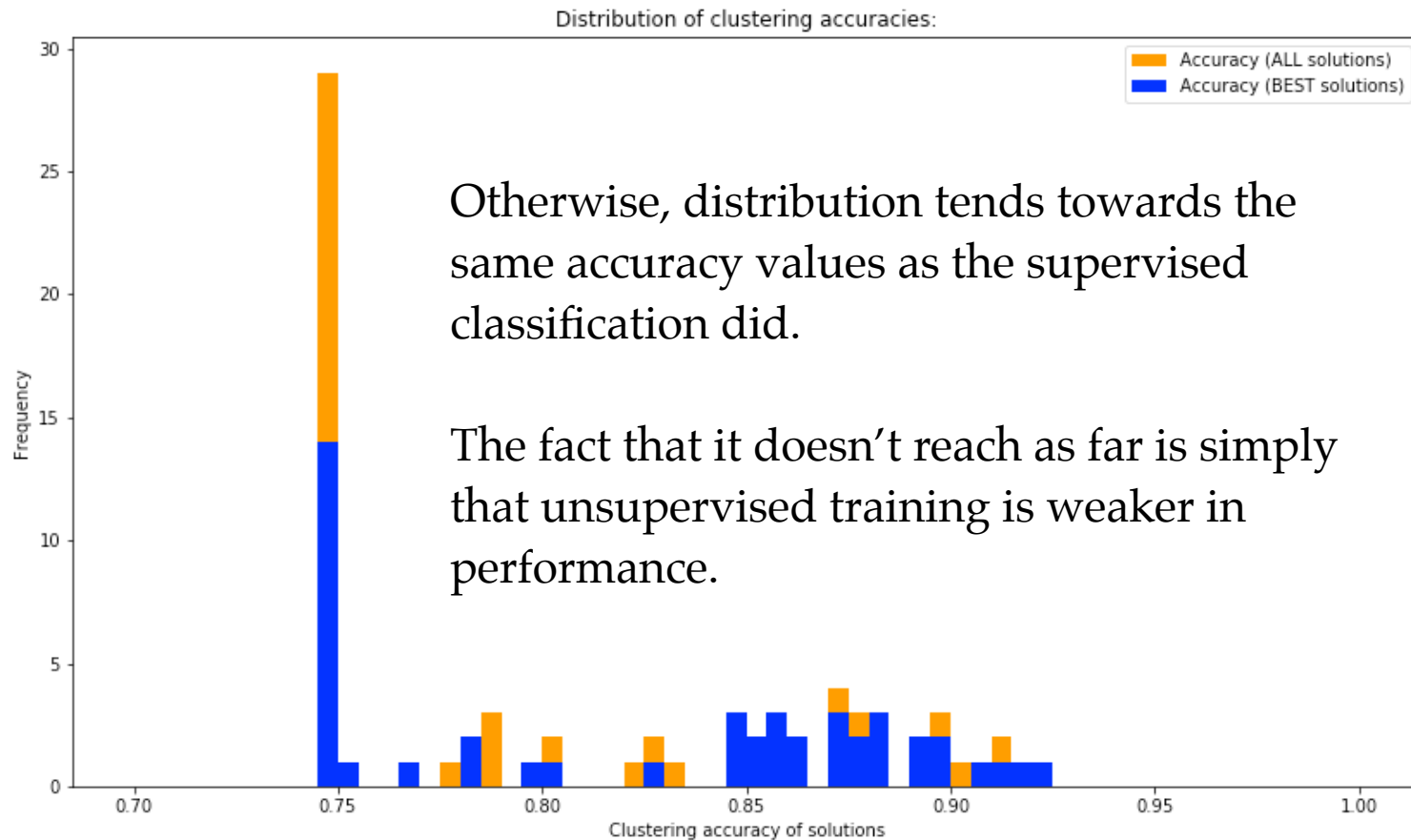
# Clustering accuracy distribution

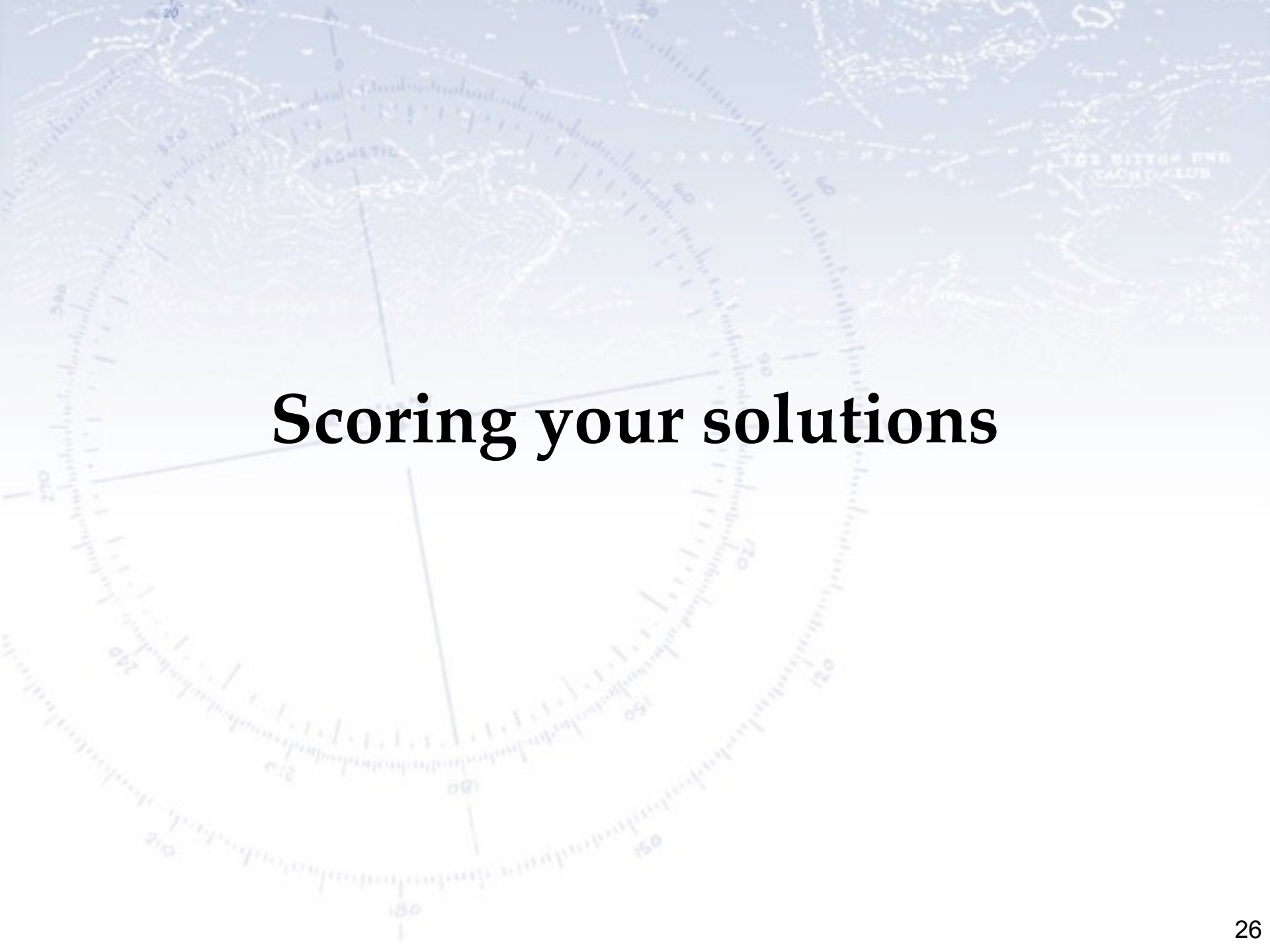
The accuracy of the clustering (when assigned either electron or not) was:



# Clustering accuracy distribution

The accuracy of the clustering (when assigned either electron or not) was:





# Scoring your solutions

# How do we grade your projects?

## Final Score:

You submitted a full solution, from which you get:

65 points

Your choice of methods based on your description was scored as follows [0,6]:

Your solutions entailed Nalgo different algorithms, which gives a score of [0,6]:

Your variable choice was scored  $8 \times (\text{Sum YourVarFreq} / \text{Sum TopVarFreq})$  [0,8]:

Your performance was for:

Classification:  $-\log(\text{CrossEntropy} - 0.14)$  [0,4+]:

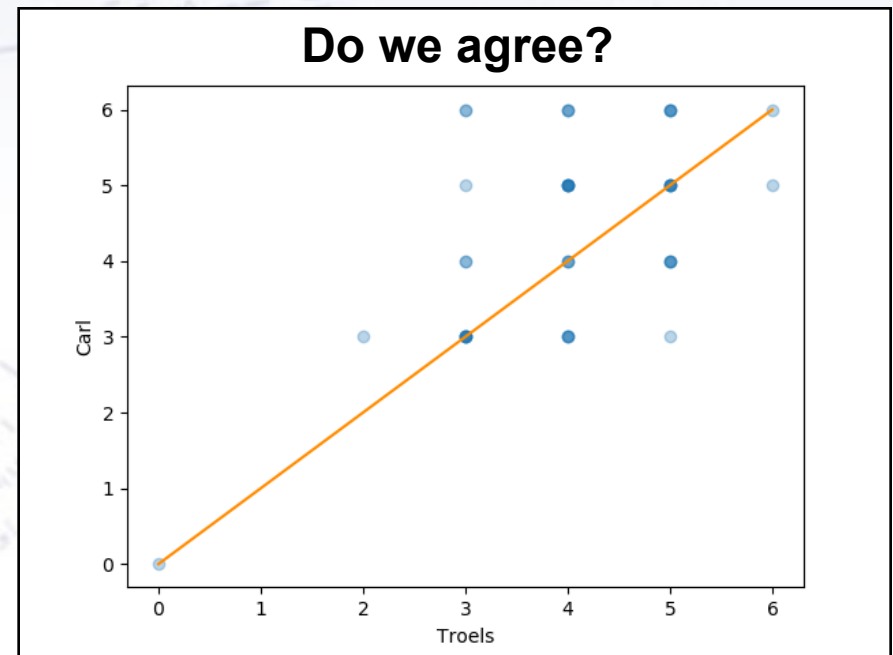
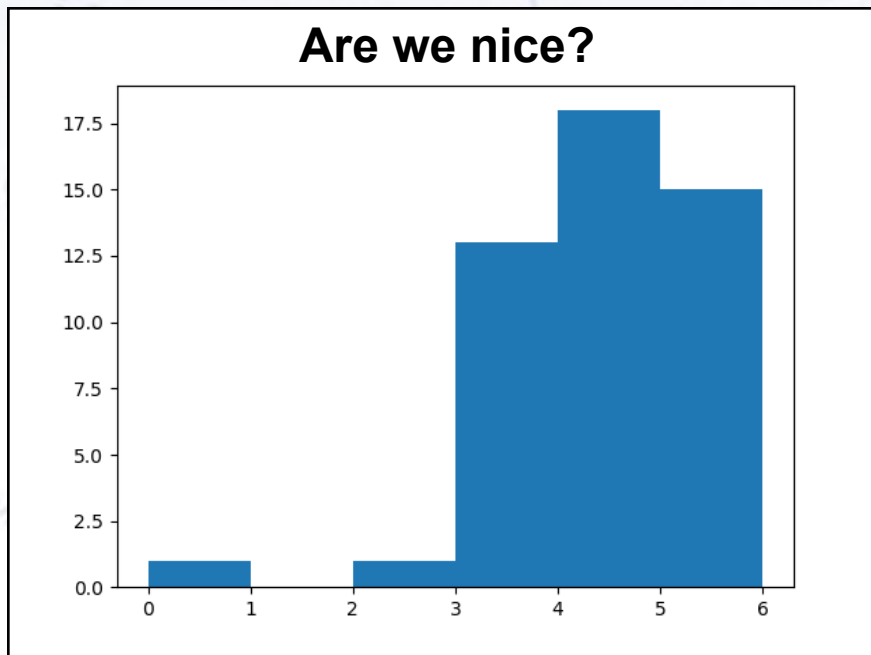
Regression:  $-\log((\text{MAE}((E-T)/T) - 8000)/8000)$  [0,4+]:

Thus your total number of points was:

N points

# Your description reports

Carl and I read through your descriptions, and did a manual scoring (the only) based on choice of algorithms, hyperparameter optimisation, and data division (e.g. cross validation). Each yielded a score of 0-2, giving a total score of 0-6 points.



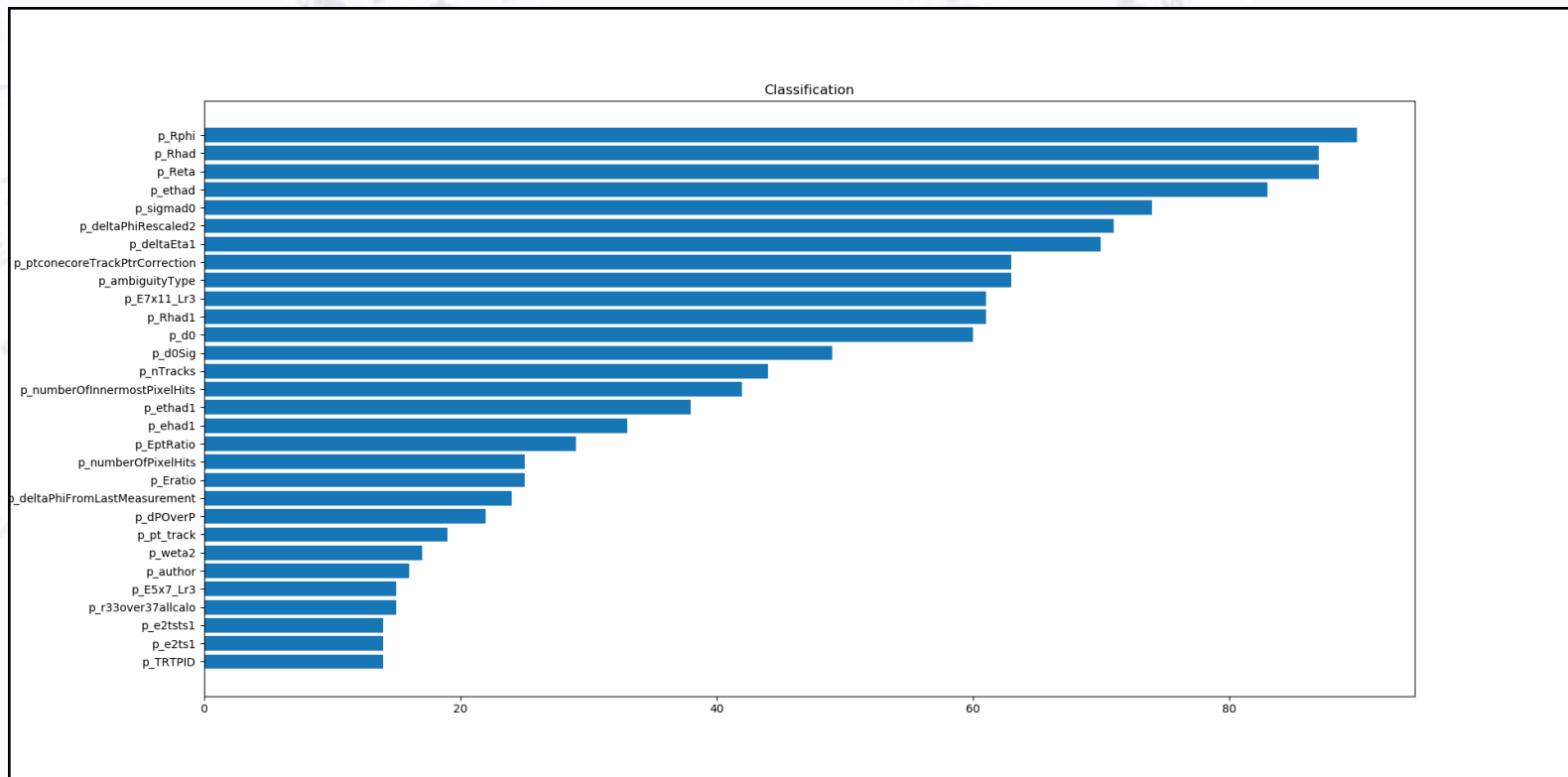
As you can see, we were generally satisfied. The descriptions were short and to the point, and give some insight into your line of thinking and working.

# Your variable choice

Assuming, that the variable frequency reflected the actual ranking very well, your variable choice was scored as follows:

$$8 \times \left( \sum Freq(\text{Your variables}) / \sum Freq(\text{Top variables}) \right)$$

...so if you picked the top variables, you would get full points.

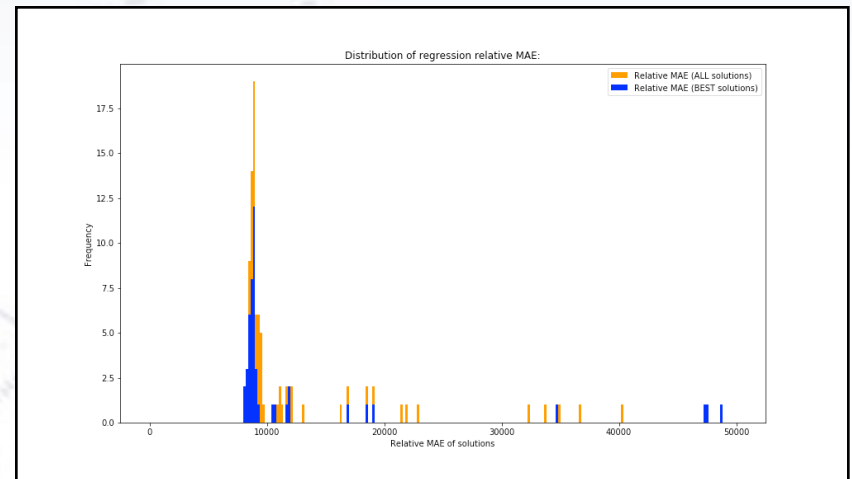
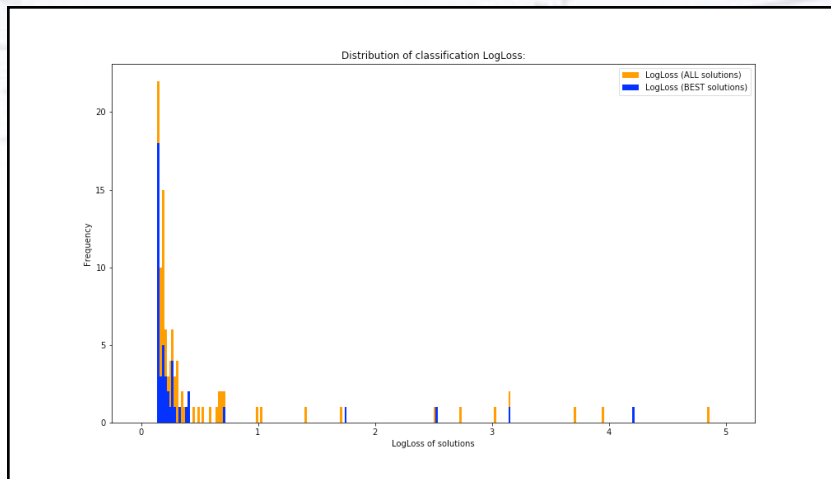




# Performance scoring

As mentioned, performance isn't everything, and we certainly didn't want it to be for the small project. Getting close to the information limit is just great.

This was reflected by using a logarithmic scoring, which turned your best key performance parameter into a score in the (open) range  $[0, 4+]$ :





**Reporting back to you**

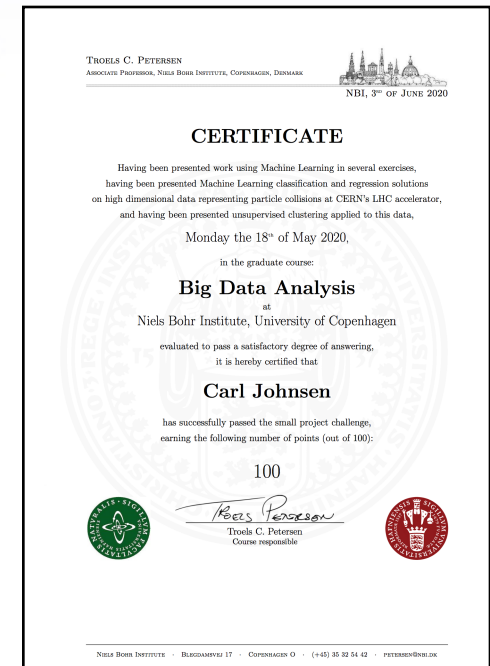
# Feedback to you

We have created a small report back to you, which consists of:

- A certificate - for you to be proud of handing in...
- A summary - for you to know how you did...
- A solution scoring with key numbers and illustrations - for you to understand how your model performed.

These are (hopefully) being mailed to you by Carl and Zoe right now. Please sit down after class and look through them.

Also, don't hesitate to discuss them with your peers. Perhaps you have already done this (great), **but this feedback and reflection is the process through which you learn the most...** please use it.



# Classification report

By now you should know what all the different plots and number are...

The solution gave the following metrics:

Metric	Equation	Value
Accuracy	<code>sklearn.metrics.accuracy_score</code>	0.940735
AUC	<code>sklearn.metrics.auc</code>	0.976952
Cross entropy	<code>sklearn.metrics.log_loss</code>	0.153488

The solution produced the following plots:

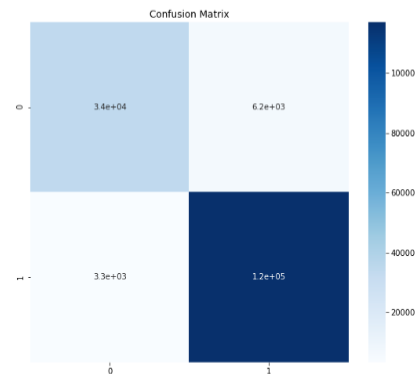
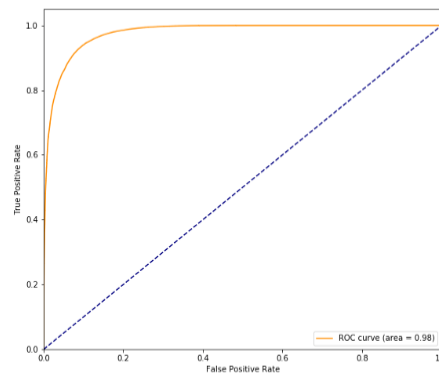


Figure 1: **Left:** ROC curve for the tensorflow2 implementation. The orange curve should be as close to the upper left corner as possible. **Right:** Confusion matrix for the tensorflow2 implementation. The diagonal squares ((0,0) and (1,1)) should have the higher values.

# Regression report

The solution gave the following metrics:

Metric	Equation	Value
MAE - Absolute	<code>sklearn.metrics.mean_absolute_error</code>	6744.0340
MAE - Relative	$\sum \frac{ y_p - y_t }{y_t}$	9019.6753
RMS	$\sqrt{\text{mean}((y_p - y_t)^2)}$	14344.1986
RMS 98th percentile	$\sqrt{\text{mean}((y_p - y_t)^2)}$	9011.5420
RMS 90th percentile	$\sqrt{\text{mean}((y_p - y_t)^2)}$	5818.4660
RMS 70th percentile	$\sqrt{\text{mean}((y_p - y_t)^2)}$	4338.8145

The solution produced the following plots:

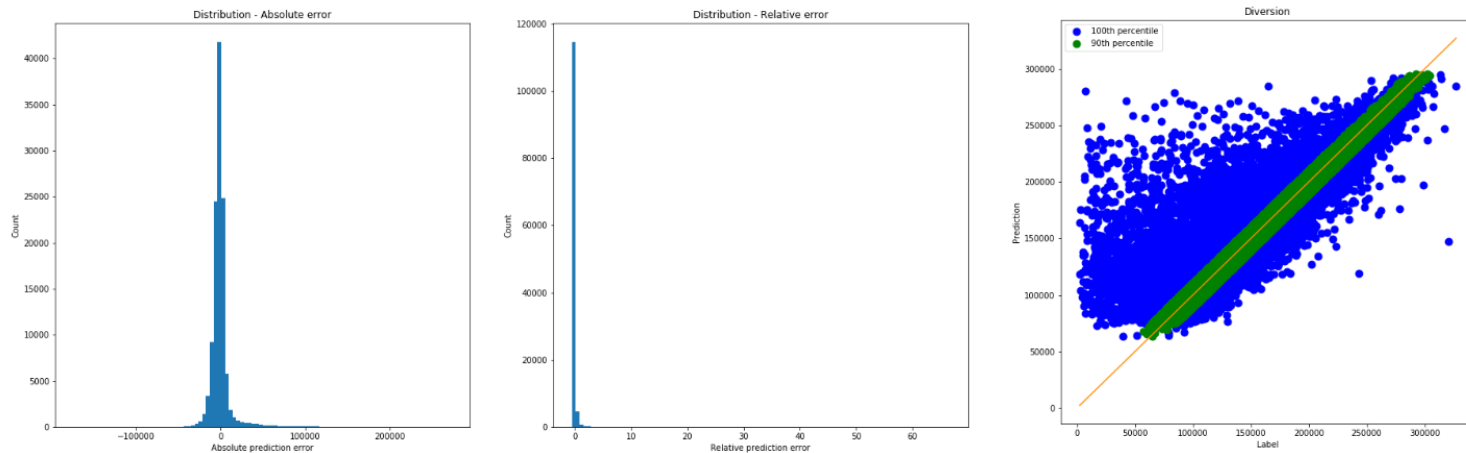


Figure 2: **Upper:** Distribution plots for the xgboost1 implementation. The plots are for absolute error (*Left*) and relative error (*Right*). Both plots should have a tall narrow curve, centered around 0. **Lower:** Diversion plot for the xgboost1 implementation. The dots should be scattered close to the line - especially for the 90th percentile.



# Clustering report

The clustering report is necessarily not very detailed, as unsupervised learning carries a great deal of uncertainty on what you're doing.

However, remember the remark by Alexander Nielsen about t-SNE, but applied more generally:

“I always start by throwing a clustering algorithm at data, just to see what structures turn up, if any. Even the latter result tells me something valuable for the further analysis.”

The solution produced the following metrics:

Metric	Equation	Value
Accuracy	<code>sklearn.metrics.accuracy_score</code>	0.8128

To compute the accuracy, the following mapping was used, based on the clusters resemblance to electron classification:

Cluster	0	1
is electron	1	0

The solution provided the following plot:

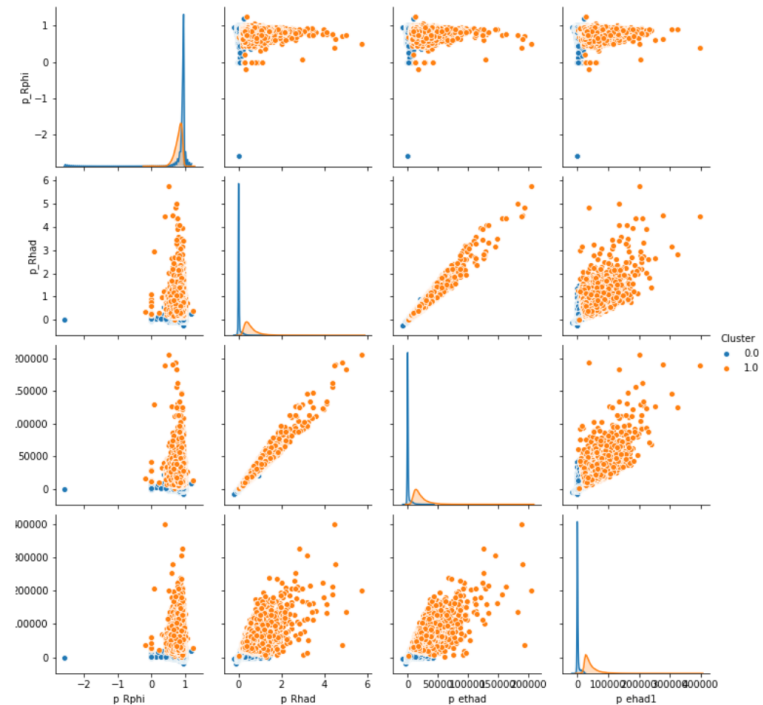
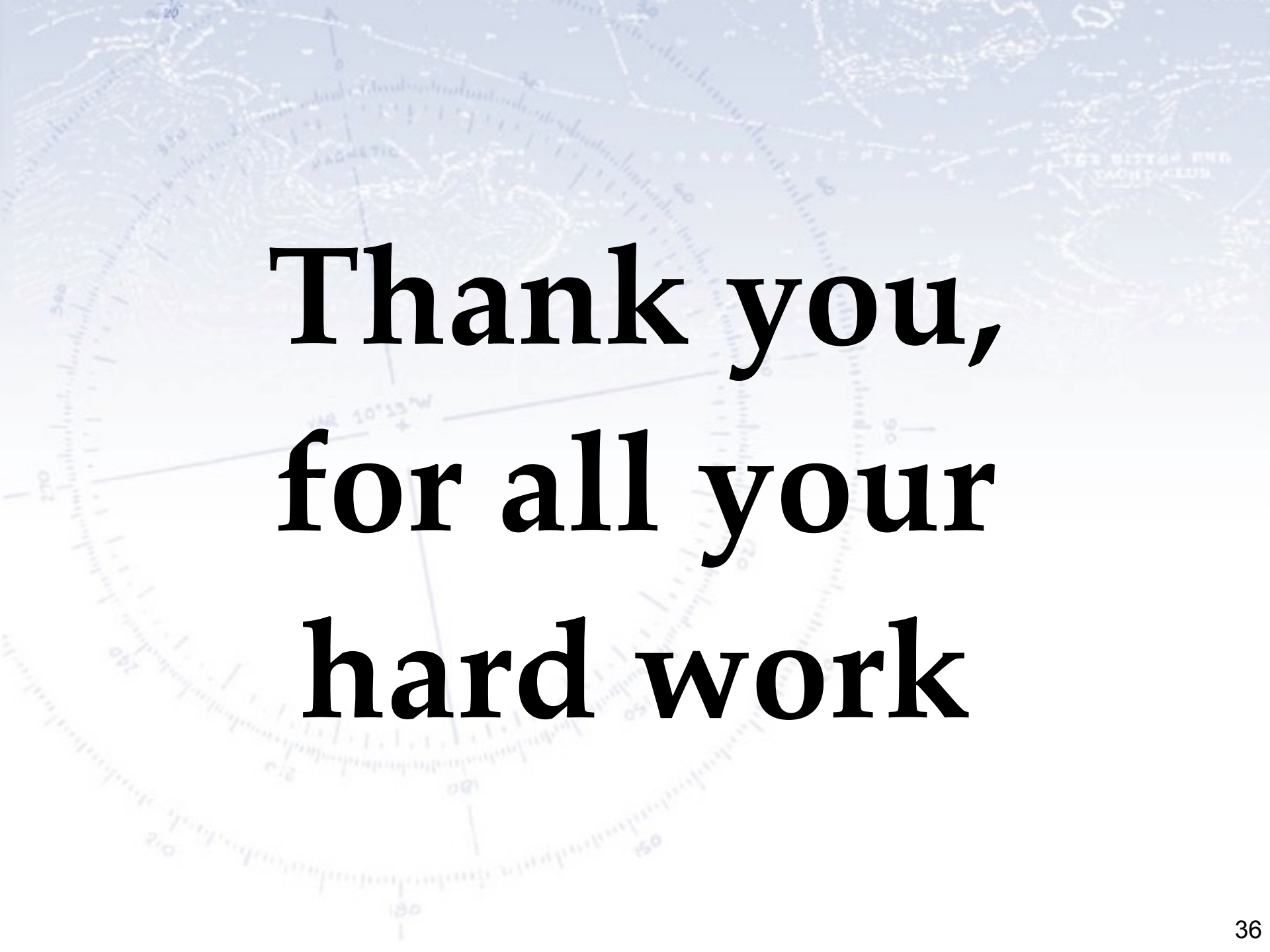


Figure 3: Pairplot for the scikitmeans1 implementation. The variables chosen are the top 4 most used variables for clustering. There should be a clear distinction of the clusters.



The background features a light blue nautical chart with a prominent compass rose. The chart includes various navigational elements such as depth soundings, magnetic variation lines, and a grid of latitude and longitude. The text is centered over this background.

**Thank you,  
for all your  
hard work**