Anomaly detection

With

Dynamic

Boltzmann machines

LUNDS UNIVERSITET
Lunds Tekniska Högskola

IBM

Hi, I’m Christoffer.

Let’s talk anomaly detection!

The agenda

1. Problem description
2. Introduction & Background
3. Methodology
4. Results
5. Conclusion

1

Studied a a combustion power plant situated in Stockholm. More specifically, one of the six boilers in the plant, Boiler 6. The goal is to detect **anomalies** in this system.

And the system is…

Combustion Powerplant

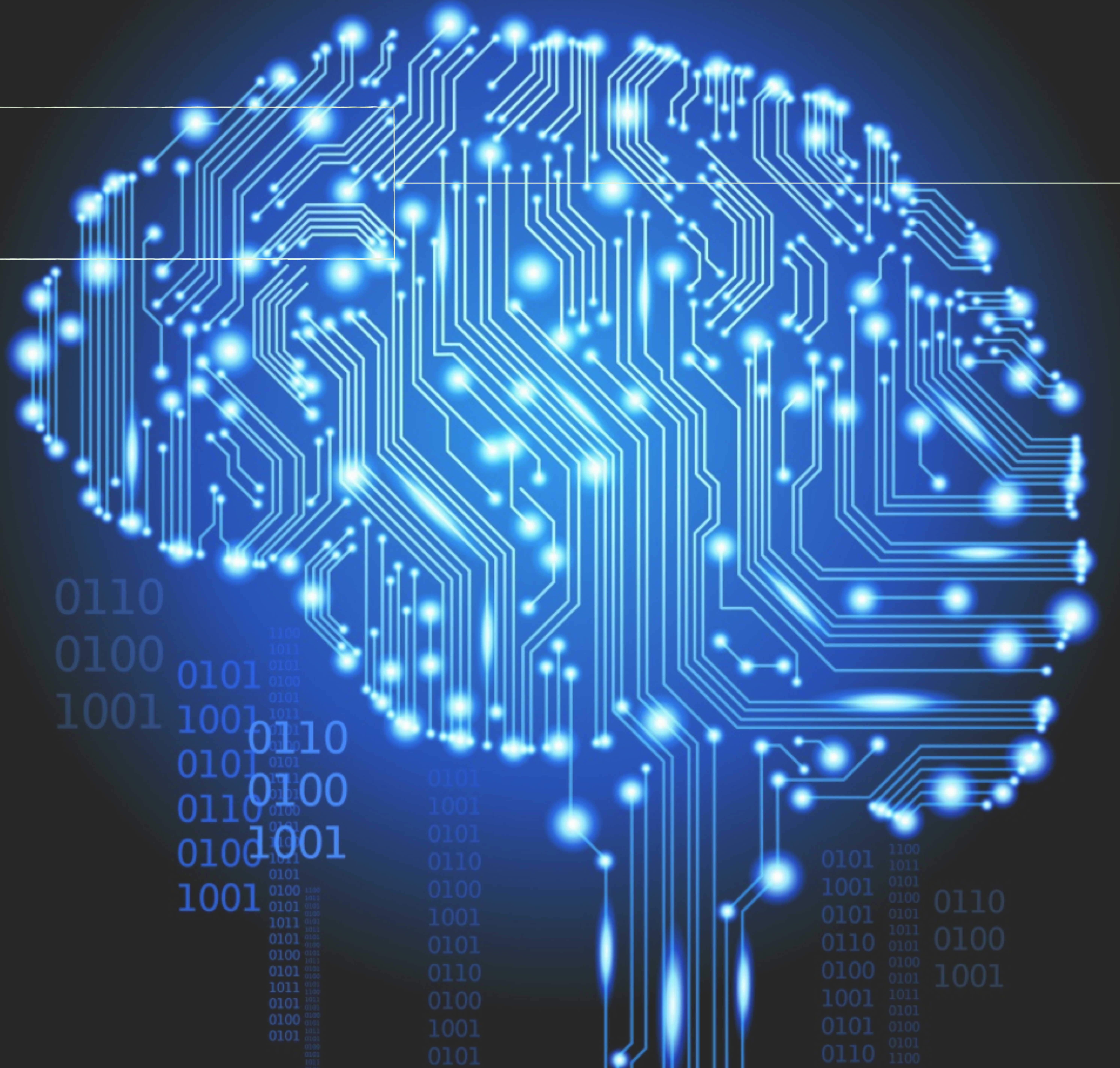
1 Boiler
3 subsystems
1000 sensors

What is anomaly detection?

”Anomaly detection are the data points in a data set that does not fit well with the rest of the data, the events or observations are classed as anomalies.”

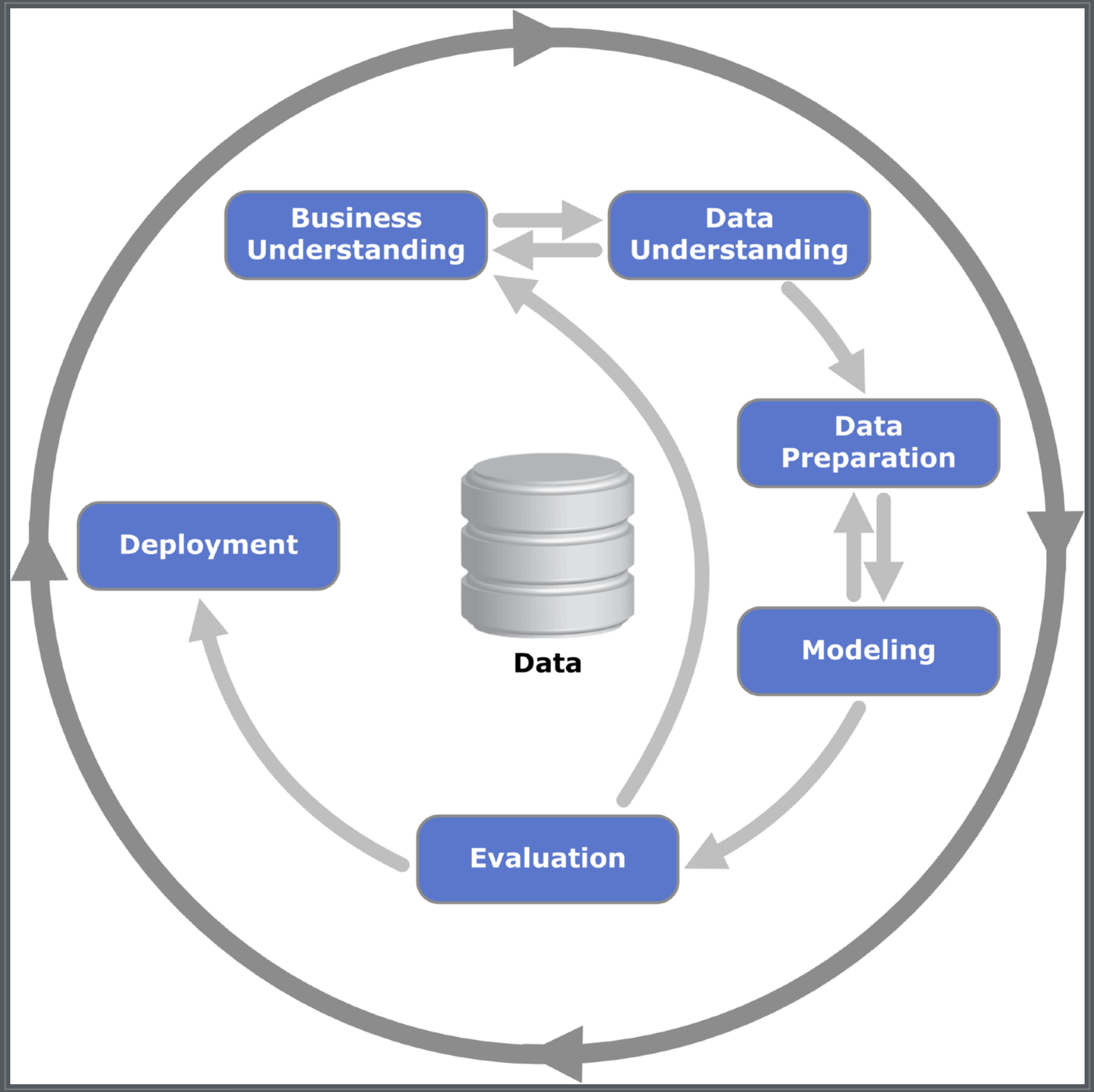
Boltzmann machines

- Type of neural network
- Energy based & stochastic model
- Increasing popularity in anomaly detection
- Too slow → Dynamic BM handles this



Methodology

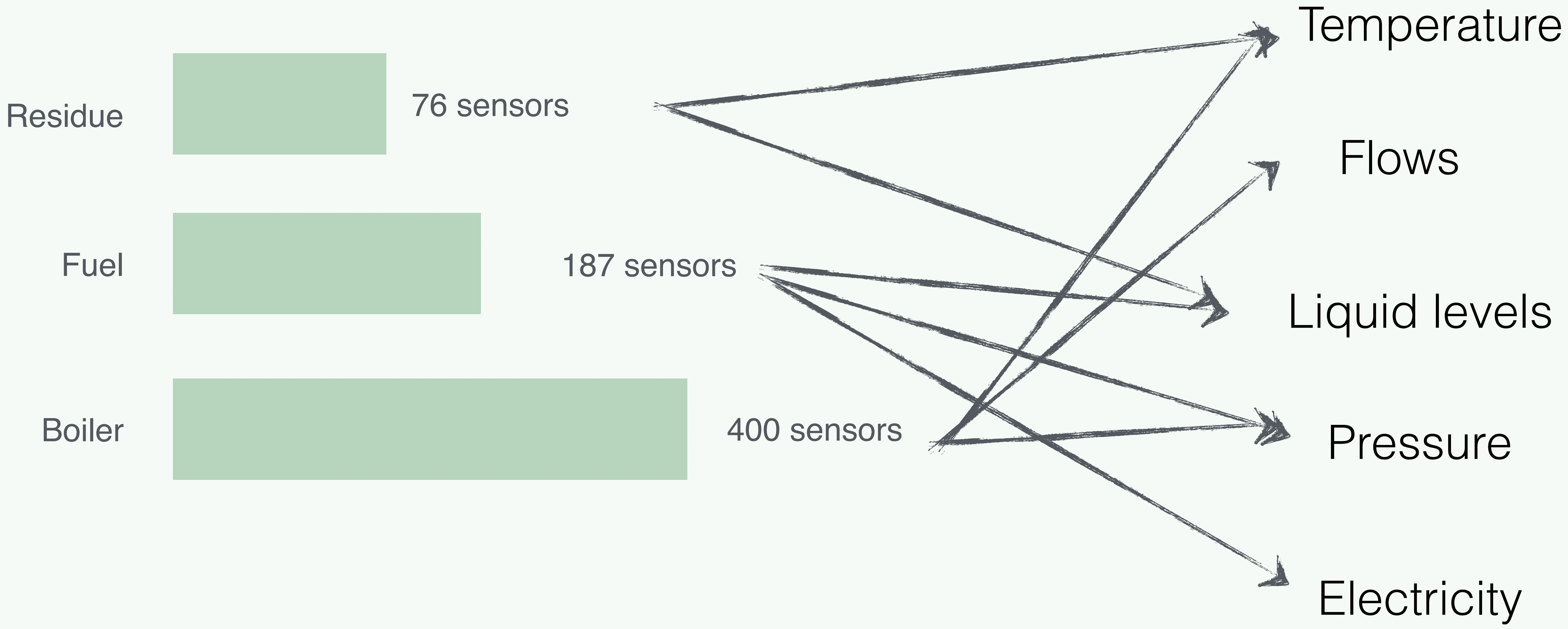
2



CRISP-DM

Cross-industry Standard Process

Subsystems



Also %

HOG.613KU201
System output

System is down

Optimal operation -
around 91 MW

Clusters

Dimensionality

Logical

Boiler - 19 clusters

Fuel - 1 clusters

Residue - 2 clusters

Table 1: Subsystems for the Boiler system and the number of variables in them

Name	# var.	Name	# var.
Ammoniak	5	Hjälpånga 40	9
Arbetsluft	4	Kylvatten	21
Avloppsånga 60	17	Manöverluft	2
Brännolja, lätt	18	Mava PN 160	45
FJV,120 C	23	MAVA S 16	9
Friblåsning	15	Pannluft	102
Hjälpånga	3	Pannvatten	7
Rökgaser	113	Sand	10
Sotningsånga	10		

Correlation

Boiler - 21 cluster

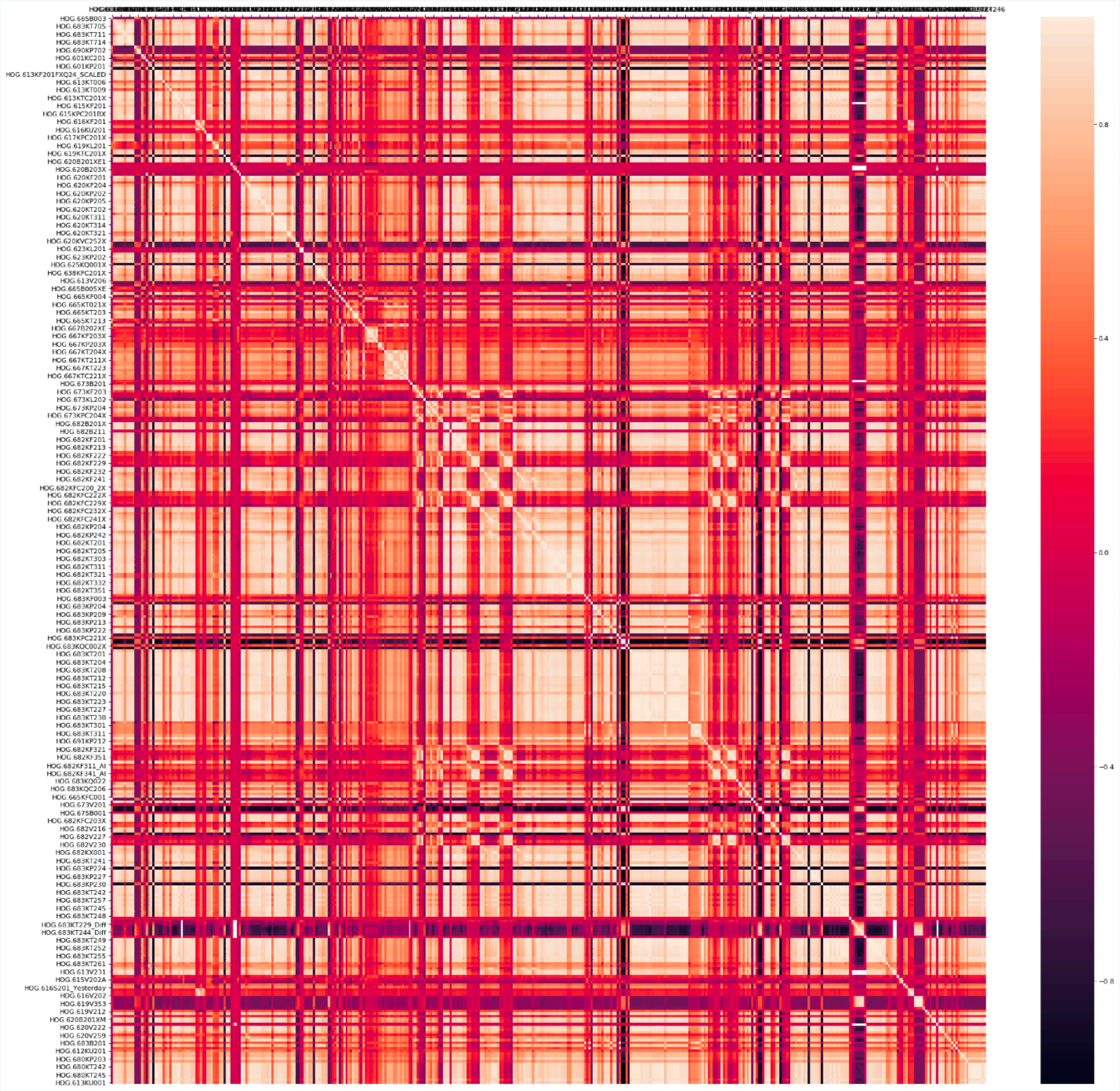
Fuel - 8 clusters

Residue - 5 clusters

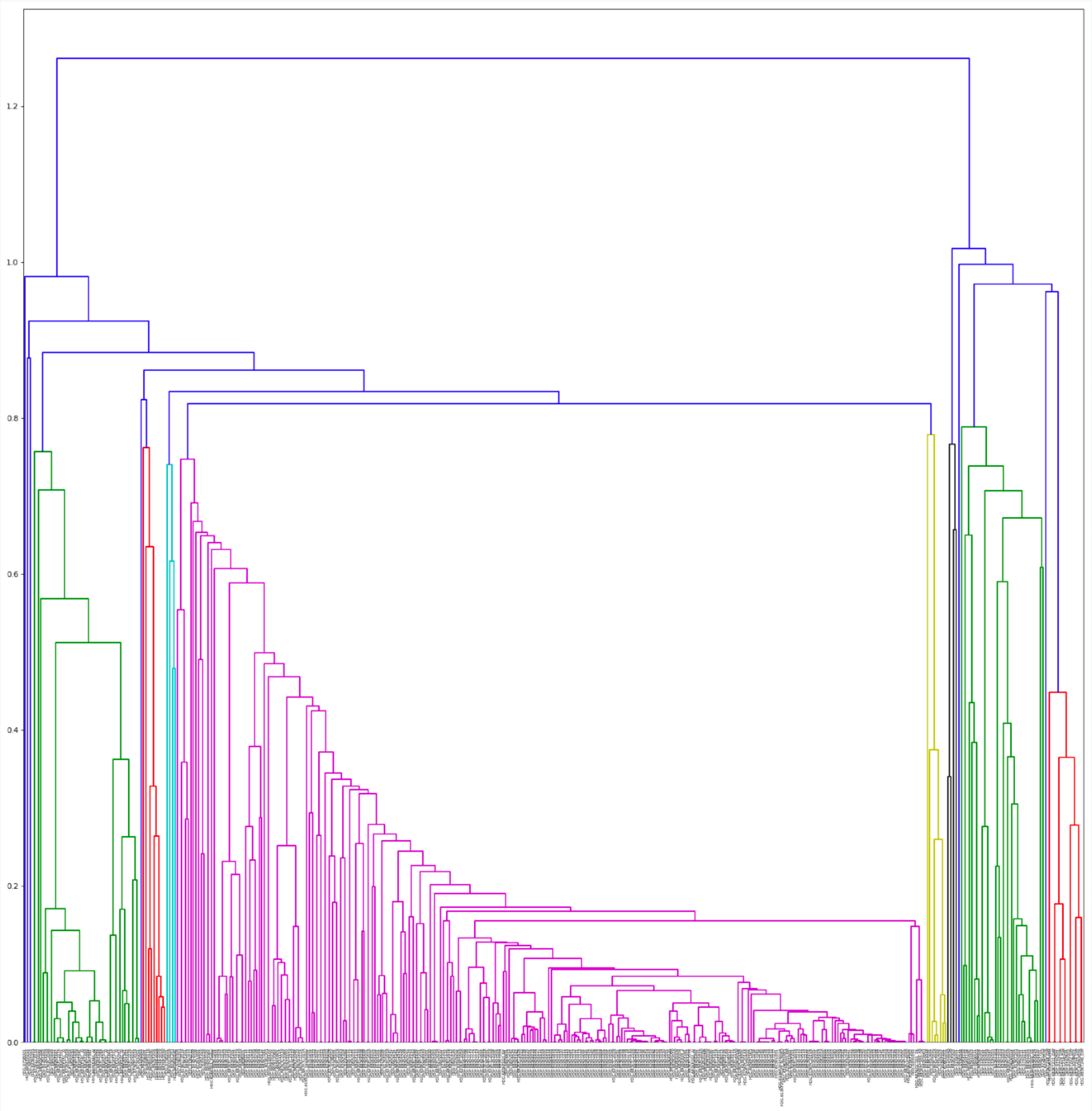
Many uni-variate cluster
where percentages

Mathematical correlations Boiler

None of the uni-variate clusters were modelled

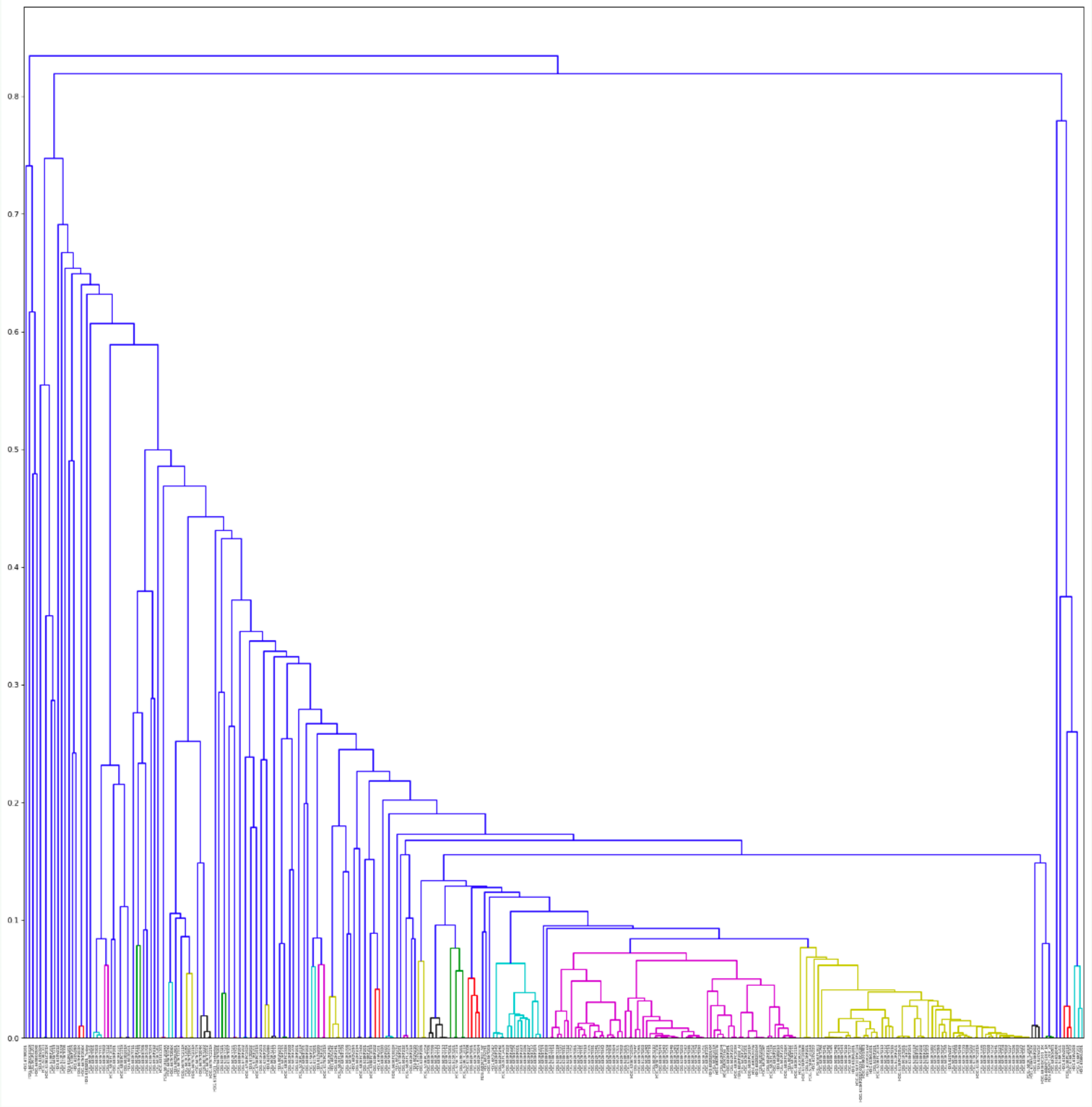


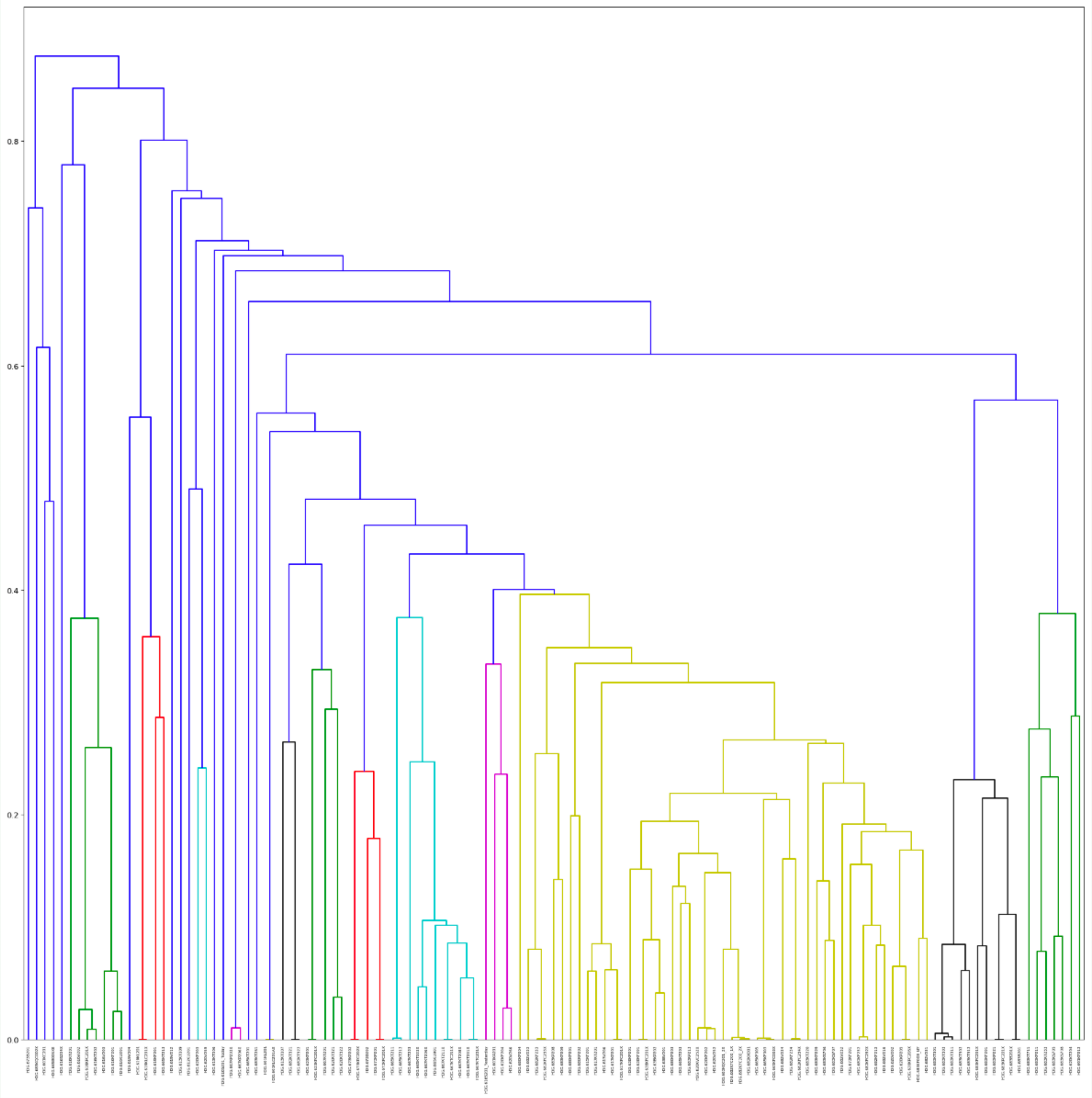
Heatmap - correlation between sensors

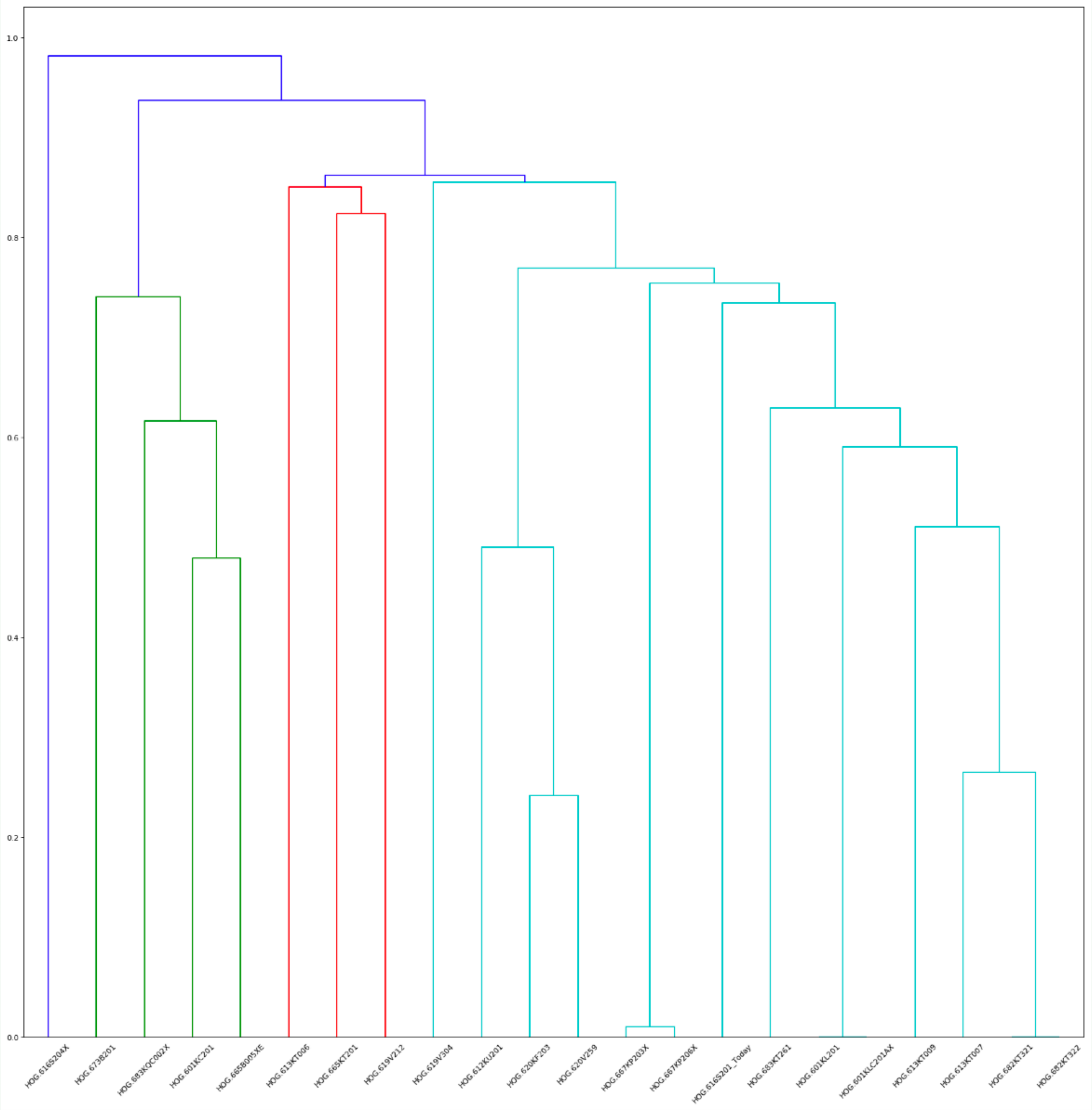


Dendrogram - clustering based on averages

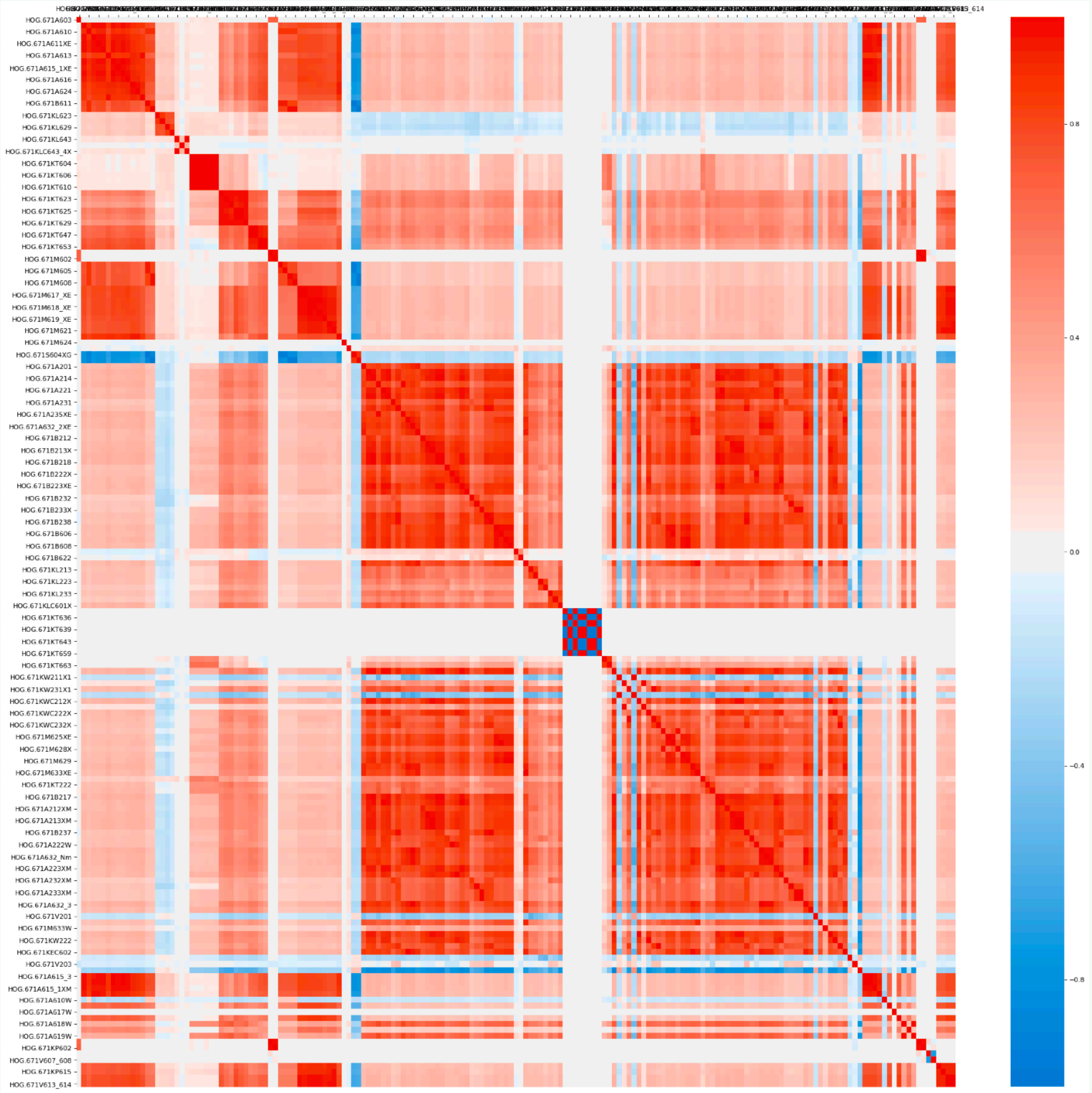
Mathematical correlations
Boiler

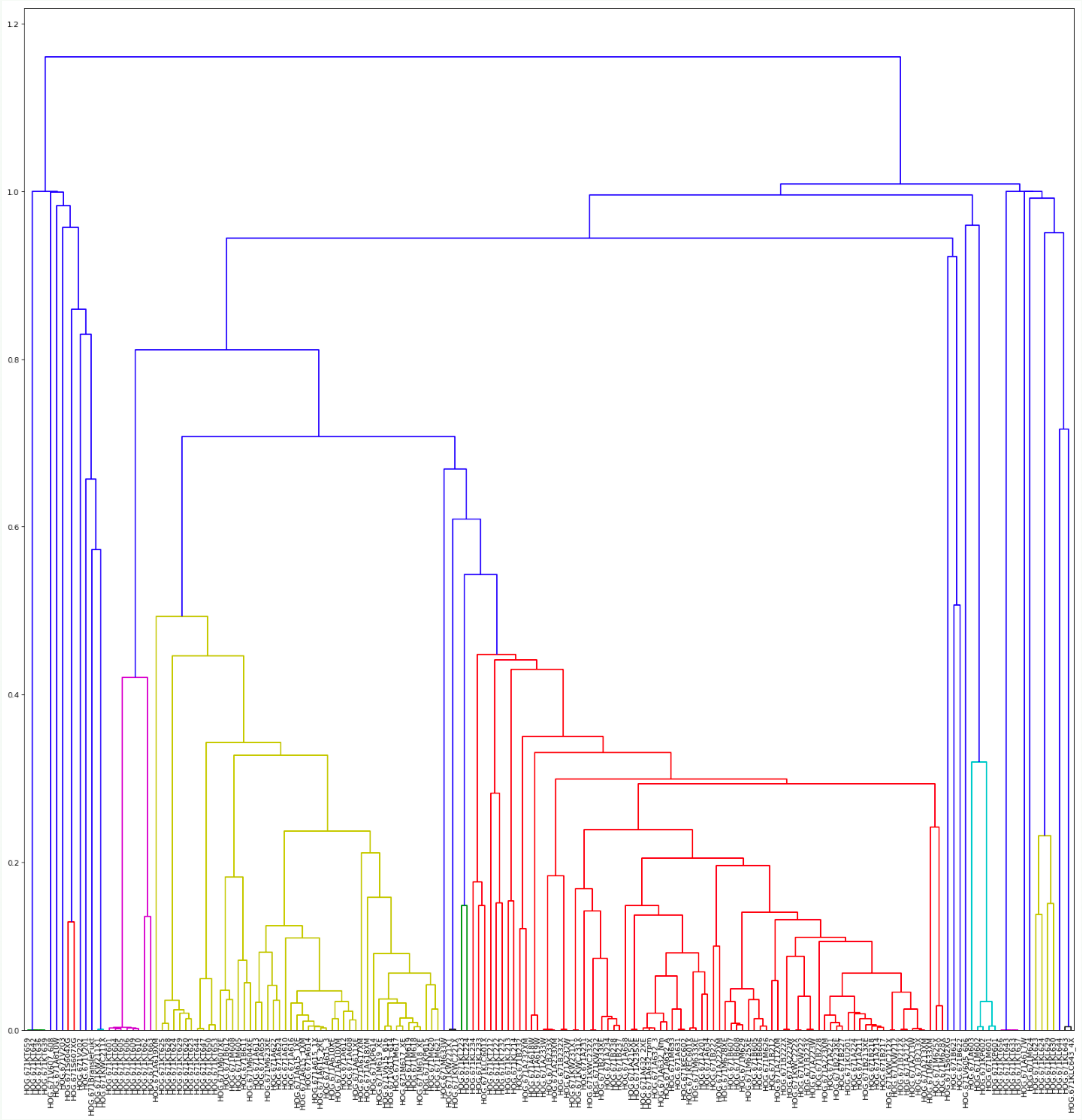




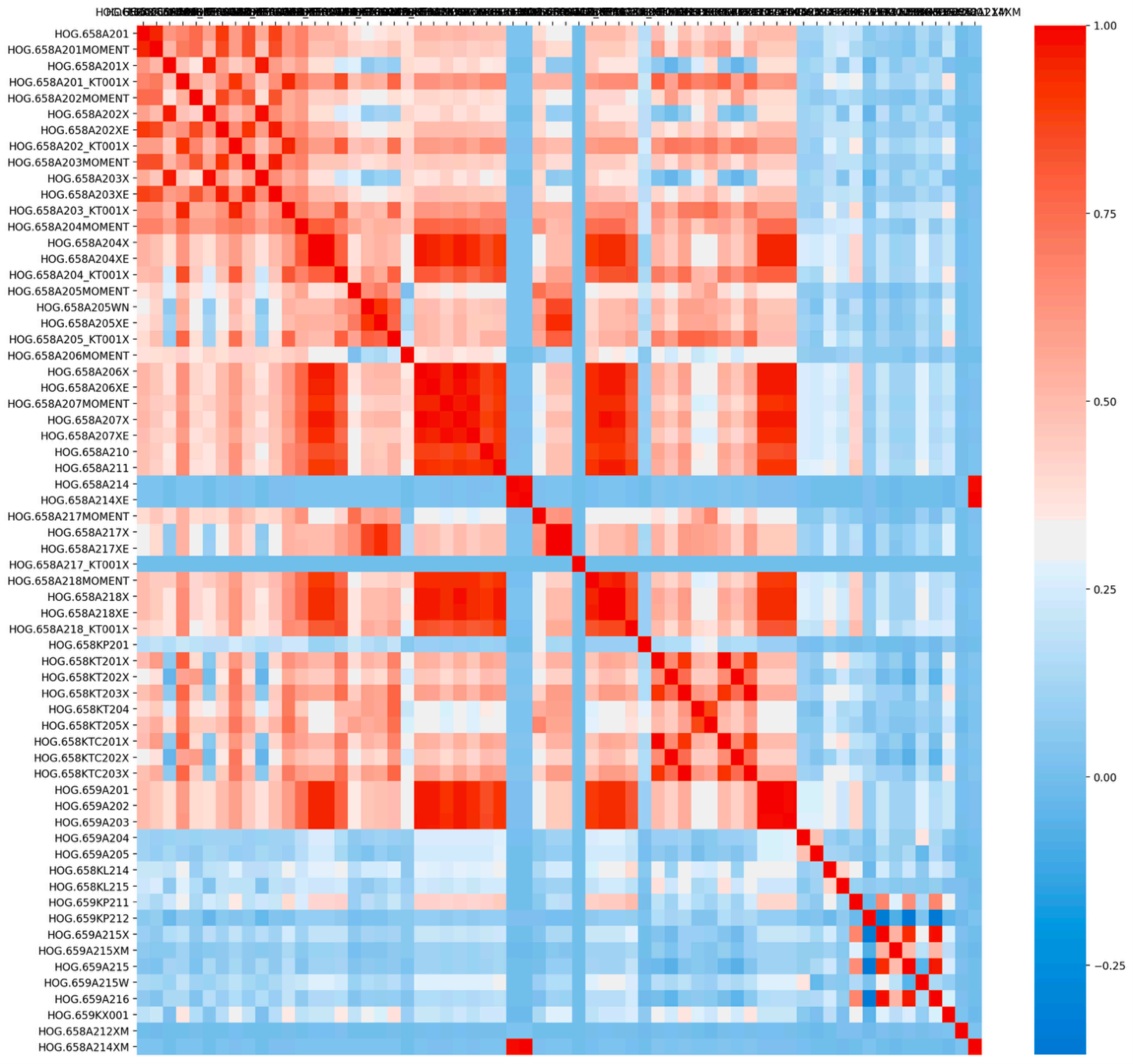


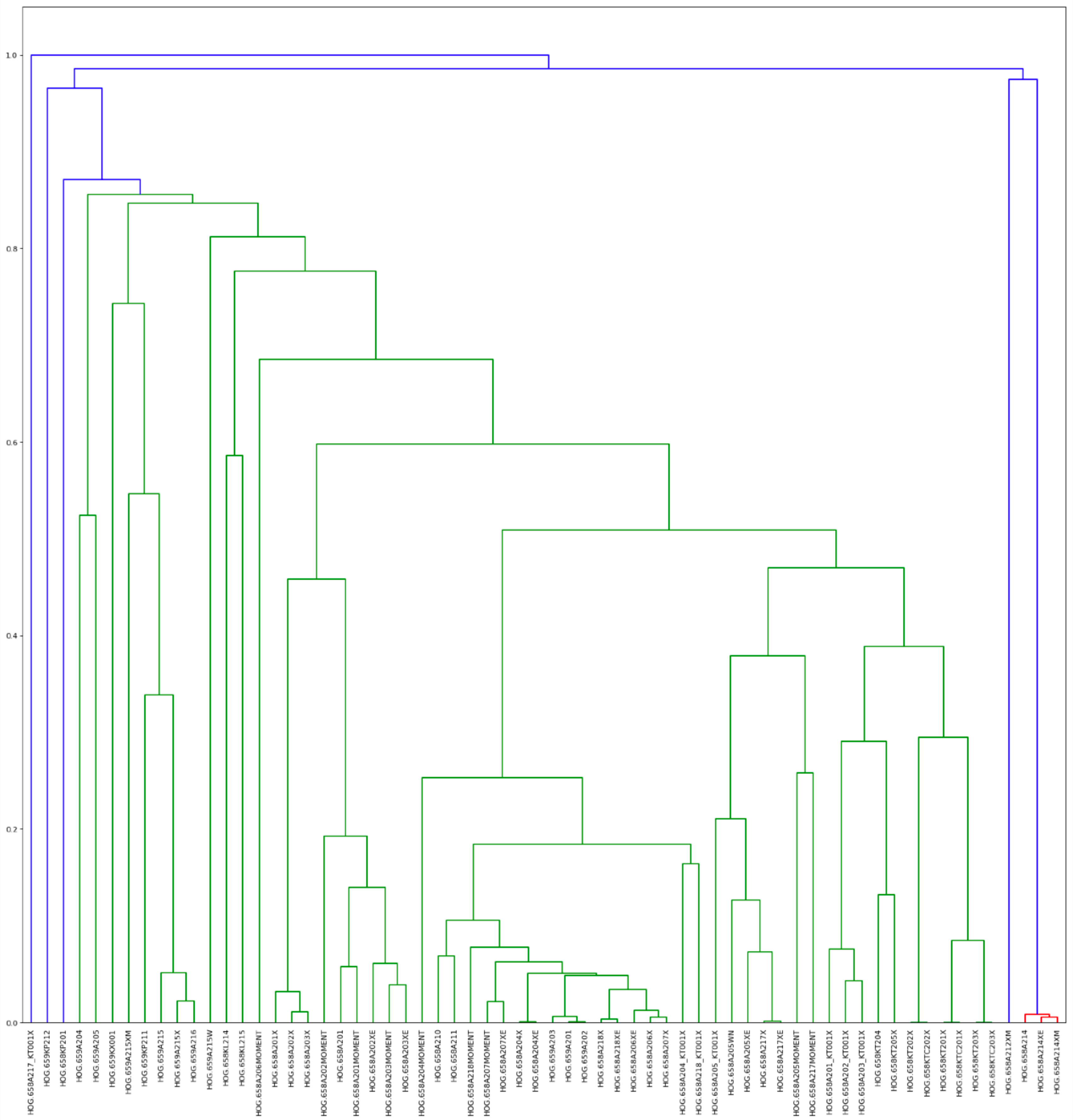
Mathematical correlations
Fuel





Mathematical correlations
Residue





Did logical and mathematical clusters match?

Math cluster	Logic cluster	Total # matches	% matches in math	% matches in logic	# math vars	# logic vars
1	Kylvatten	4	100	19	4	21
2	Pannluft	4	100	3	4	107
3	Rökgaser	6	100	5	6	114
4	Hjälpånga 40	2	40	22	5	9
4	Friblåsning	2	40	13	5	15
5	Avloppsånga 60	1	25	5	4	19
5	Pannluft	3	75	2	4	107
6	Rökgaser	10	71	9	14	114
6	Pannluft	4	28	4	14	107
7	Rökgaser	19	27	17	70	114

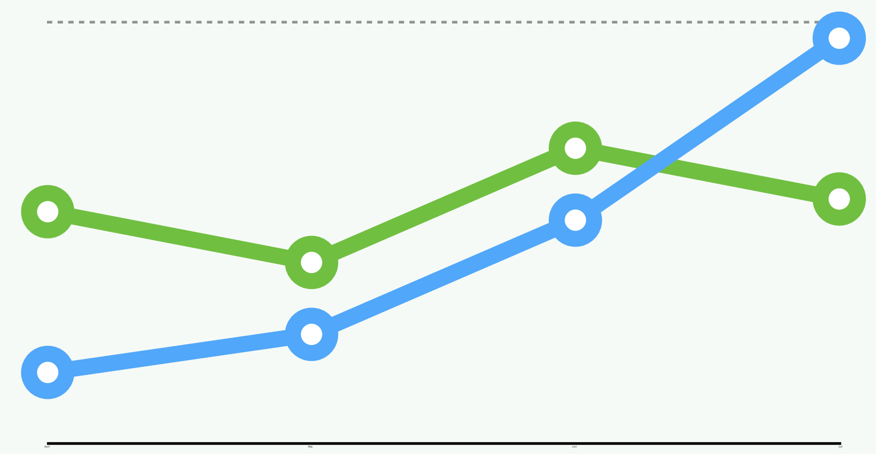
Did logical and mathematical clusters match?

Sort off

Math clusters more precise

Not as strong correlation in logical groups as believed

Modeling



One model for each cluster -
34 different models!

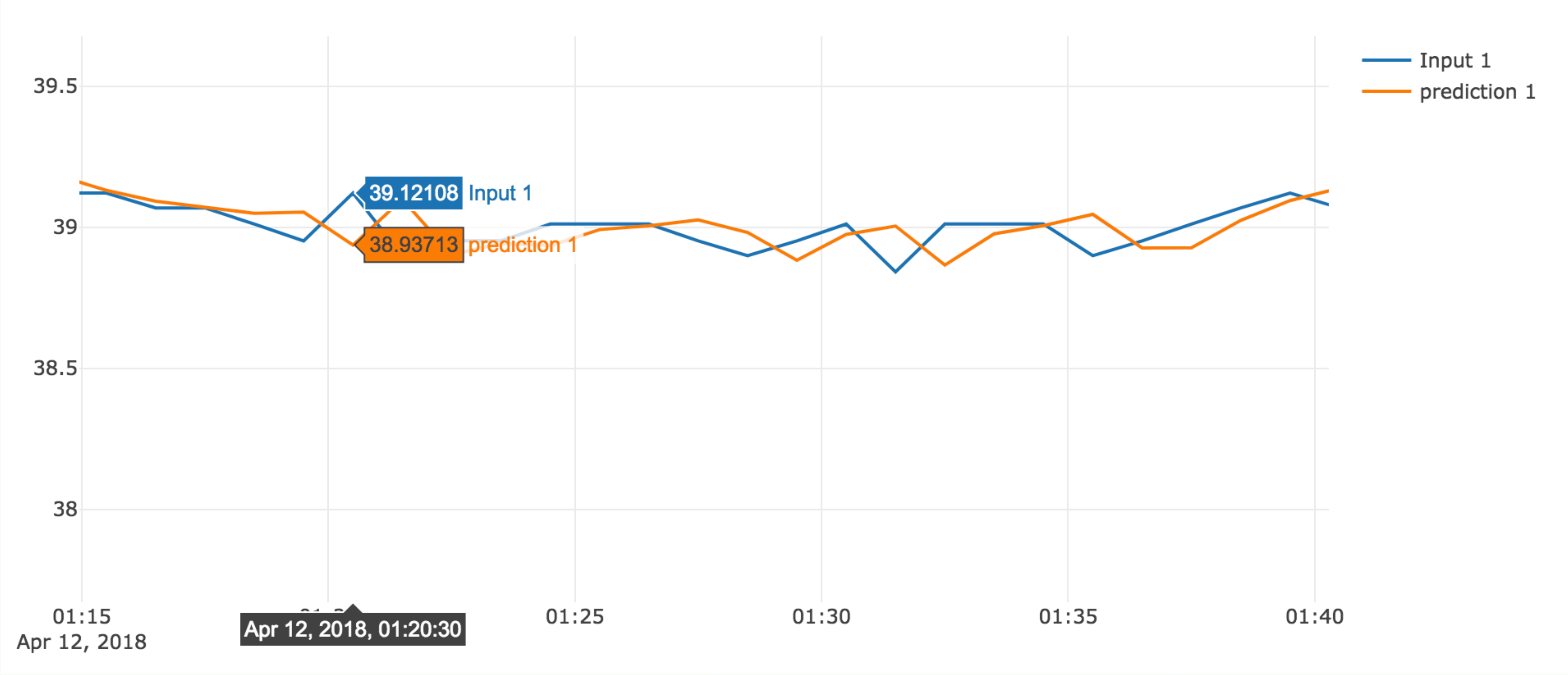
Setting the correct parameters such as learning rate, decay rate, delay and threshold

Creating a multivariate Model

Optimize parameters with RMSE

40/60 ratio on training & test data

Calculating the multivariate

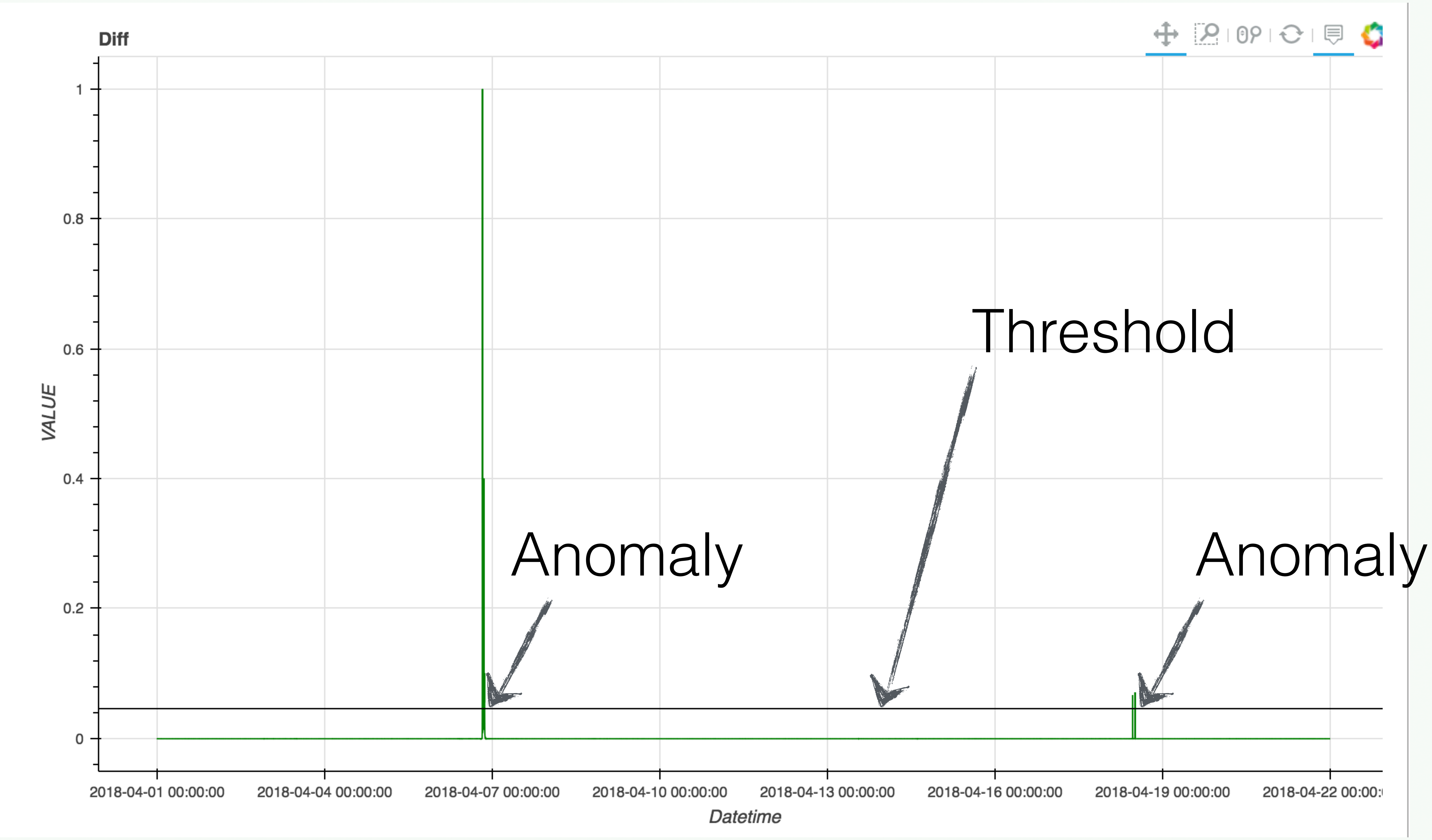


1. Take difference between input and prediction for all vectors, normalize and sum them all together

2. Calculate the Interquartile range

Issue: How to take into account different magnitudes of sensors?
Do they all contribute as much?

Determining the threshold



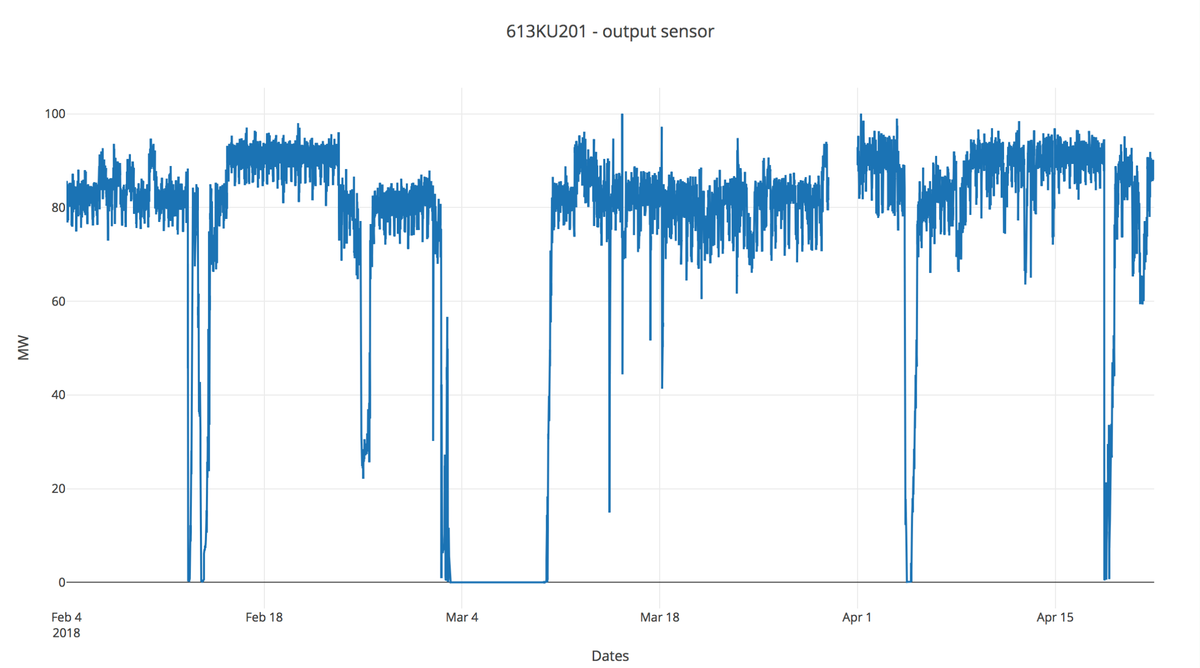
Threshold is determined by the upper interquartile range

Precision & Recall
Boiler Clusters

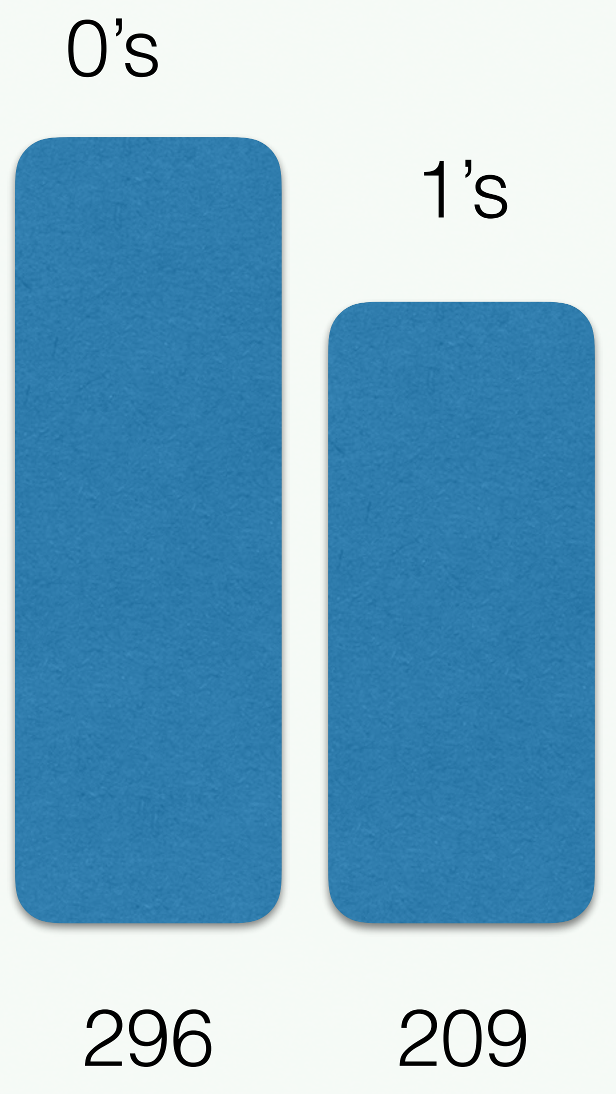
Issue: Converting to hourly resolution vs minute.

PPCR identifies the percentage of the total population that is flagged

Including all clusters give 100% PPCR



Precision: 65%
Recall: 46%
Accuracy: 55%
PPCR: 41%



209 anomaly hours during 1-22 April

Conclusion

Many open questions still, what is the best threshold method, when to use a more simple method etc.

Good results visually, hard to quantify without better labelled data.

Thank you!