ATB
Leibniz-Institut für
Agrartechnik und Bioökonomie

Data Science & Machine Learning
RESULTS

Dr. Julian Adolphs
Department Data Science

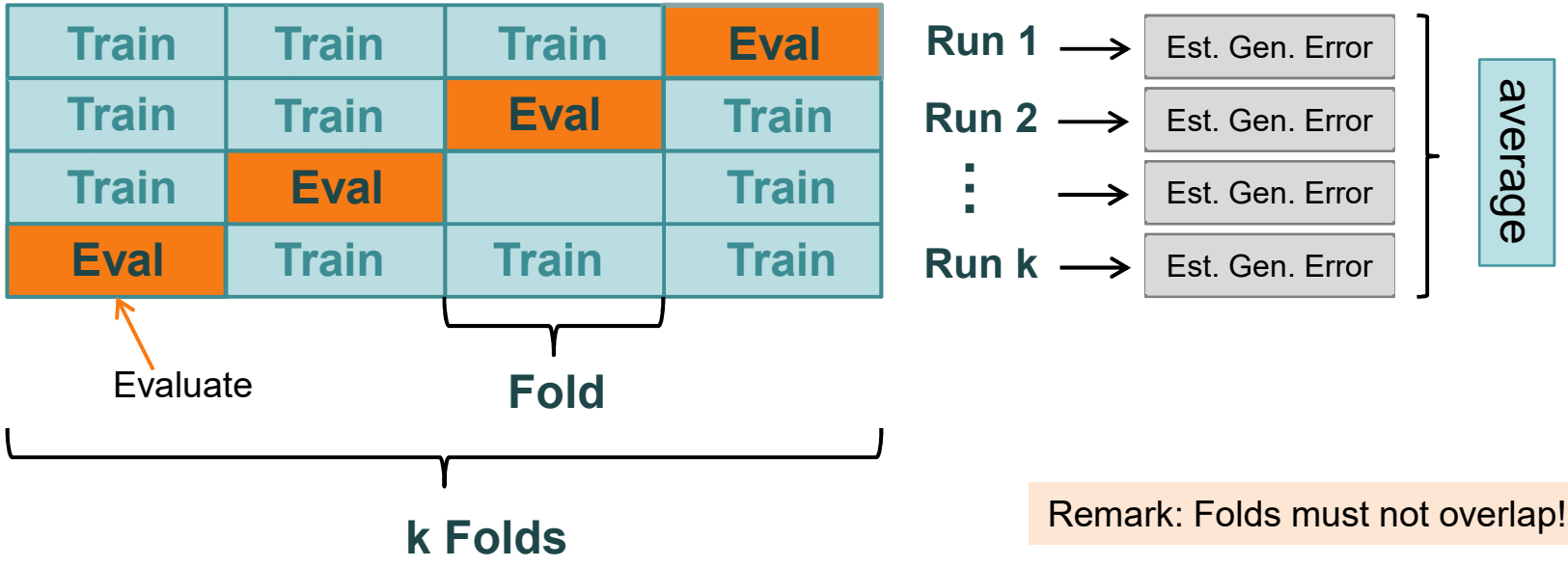
Leibniz
Association

K-Fold Cross Validation

The *Holdout Method* is a bit **wasteful** use of data. **K-Fold** is more efficient:

On each run the procedure of the former slide is performed.

The Estimated Generalization Error is the average of k runs.

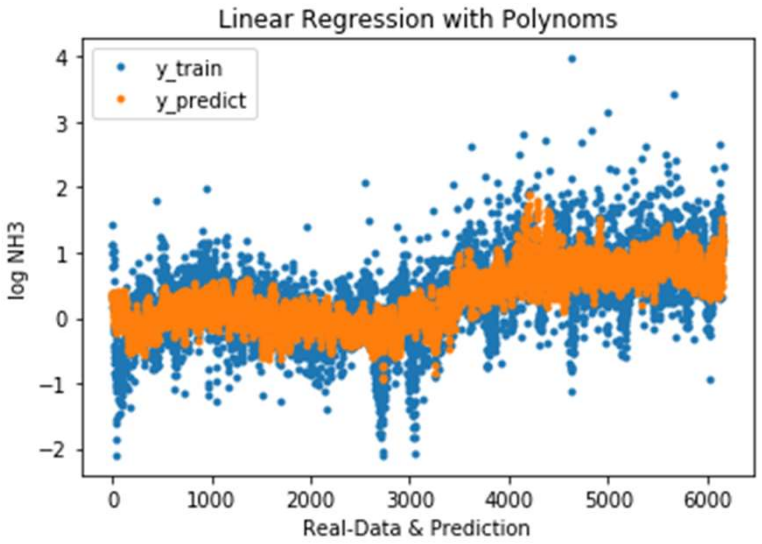


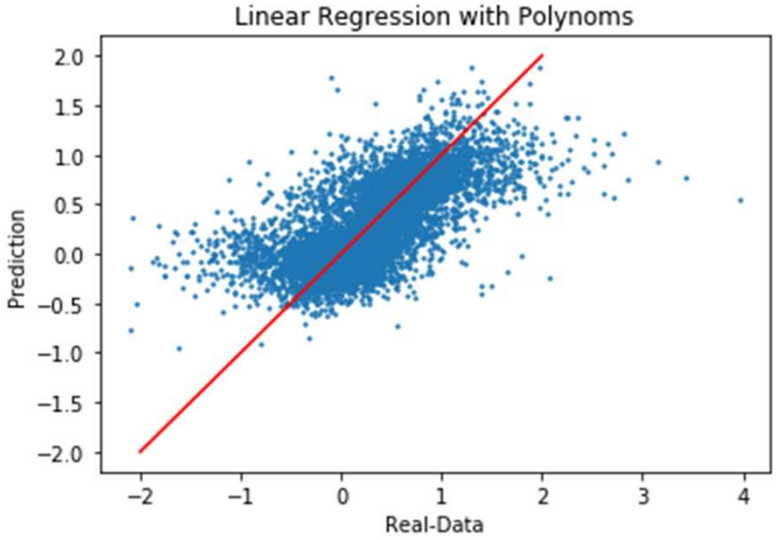
Remark: Folds must not overlap!

NH3-Measuring Method 1

Cross-Validation gives estimates for the quality of a method
(**M**ean **A**bsolute **E**rror, smaller number is better):

Linear Regression,	MAE-Score:	0.351
Linear Regression Polynom,	**MAE-Score:**	**0.332** ←
Random Forest,	**MAE-Score:**	**0.339**
Support Vector Machine,	**MAE-Score:**	**0.337**
Artificial Neural Net,	MAE-Score:	0.400

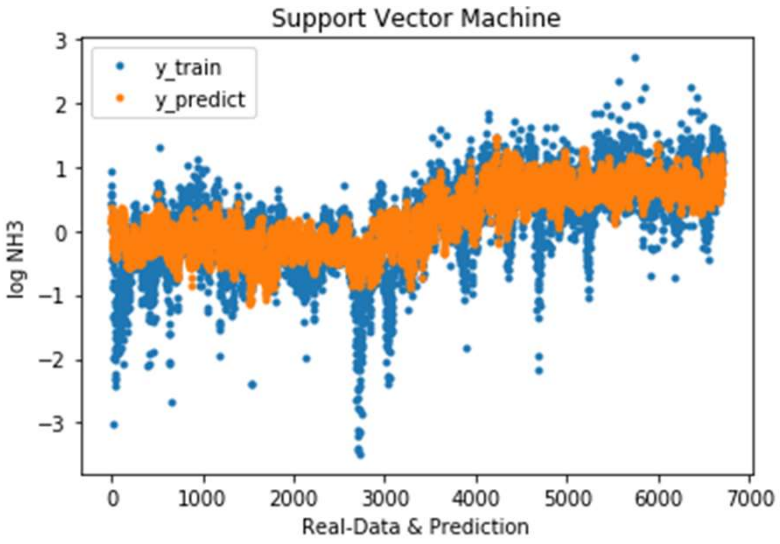


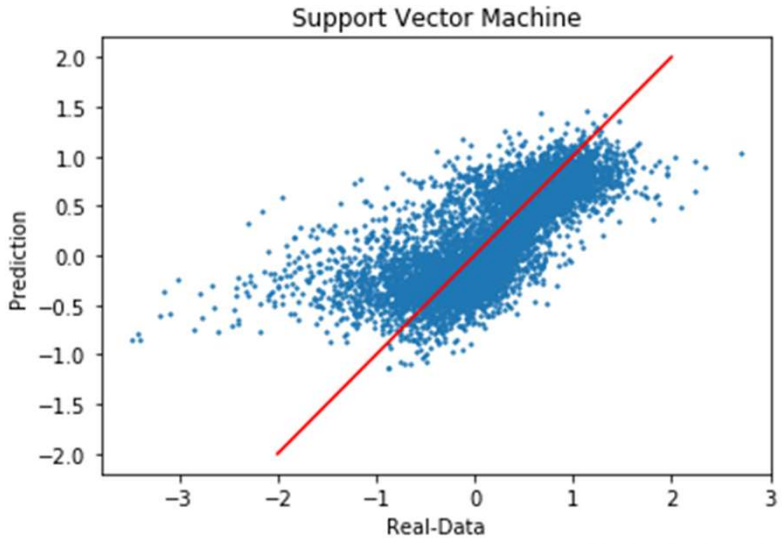


NH3-Measuring Method 4

Cross-Validation gives estimates for the quality of a method
(**M**ean **A**bsolute **E**rror, smaller number is better):

Linear Regression,	MAE-Score:	0.356
Linear Regression Polynom,	**MAE-Score:**	**0.330**
Random Forest,	**MAE-Score:**	**0.336**
Support Vector Machine,	**MAE-Score:**	**0.328** ←
Artificial Neural Net,	MAE-Score:	0.363





Reference Value of Emission in kg for Measuring Period (10 months) and per Livestock Unit

Exp.-Method 1:

Sum over measured values: **10.31 kg**
Sum over measured values with correction: **11.45 kg**

(correction for missing values: correction factors)

Exp.-Method 4:

Sum over measured values: **9.44 kg**

(only very few missing values, so correction not needed)

Train on 4 weeks with **restriction W, S, T, T**

Exp.-Method_1:

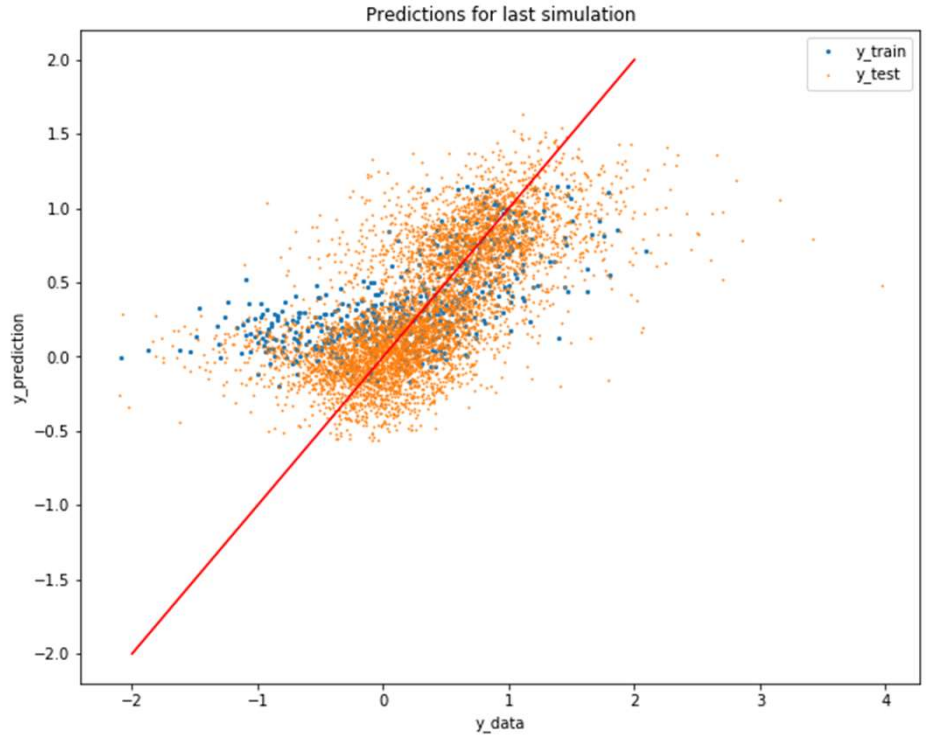
Polynomial Regression:			**Exp. Ref.**	
EF (y_train + y_test_pred), mean, std:	**9.2**	**(1.0)**	**10.3**	**(1.1)**
Extrapolation for 10 months from y_train:	10.0	(1.7)		
Extrapolation for 10 months from y_train_pred:	9.1	(1.4)		

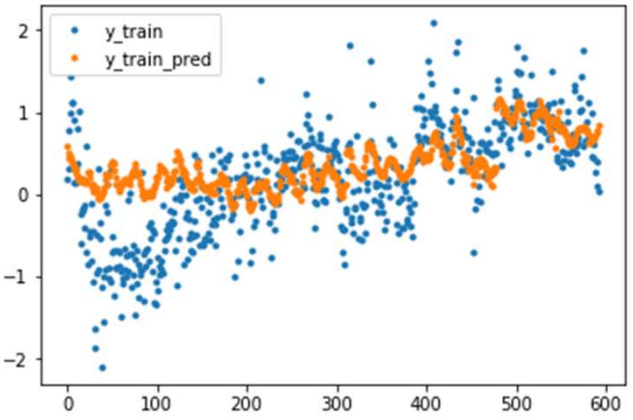
Random Forest:		
EF (y_train + y_test_pred), mean, std:	**9.7**	**(1.1)**
Extrapolation for 10 months from y_train_pred:	9.6	(1.6)
Extrapolation for 10 months from y_train:	10.0	(1.7)

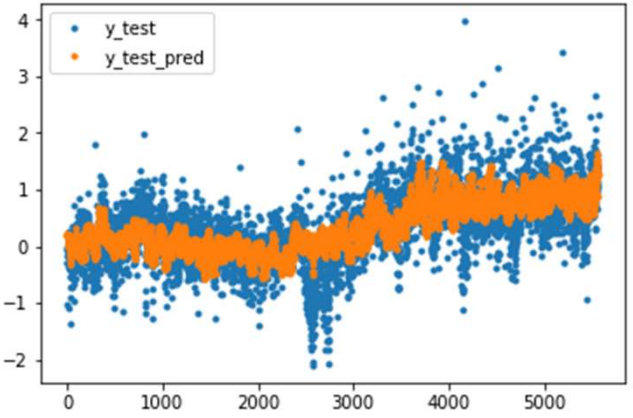
Support Vector Machine:		
EF (y_train + y_test_pred), mean, std:	**10.2**	**(7.2)**
Extrapolation for 10 months from y_train_pred:	9.2	(1.5)
Extrapolation for 10 months from y_train:	10.0	(1.7)

Units: kg per LU and per measuring period (10 months)

Prediction with Polynomial Regression (Exp.-Method 1)







Train on 4 weeks with **restriction W, S, T, T**

Exp.-Method_4:

Polynomial Regression:		**Exp. Ref.**
EF (y_train + y_test_pred), mean, std:	**8.9 (1.2)**	**9.44**
Extrapolation for 10 months from y_train:	8.8 (1.4)	
Extrapolation for 10 months from y_train_pred:	8.4 (1.3)	

Random Forest:	
EF (y_train + y_test_pred), mean, std:	**9.0 (1.2)**
Extrapolation for 10 months from y_train_pred:	8.8 (1.3)
Extrapolation for 10 months from y_train:	8.8 (1.4)

Support Vector Machine:	
EF (y_train + y_test_pred), mean, std:	**10.6 (13.1)**
Extrapolation for 10 months from y_train_pred:	8.7 (1.3)
Extrapolation for 10 months from y_train:	8.8 (1.4)

Units: kg per LU and per measuring period (10 months)

Train on 4 weeks with **restriction W, S, T, T**

Exp.-Method_4:

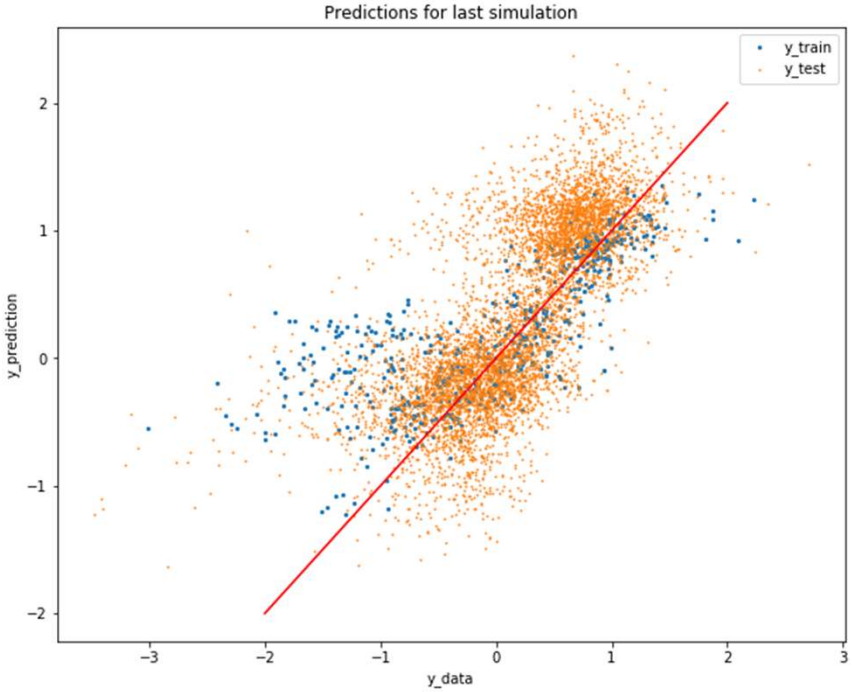
This crazy **high value** is caused by
a single week combination:
[15, 18, 20, 29] with EF = 108.3 kg (!)

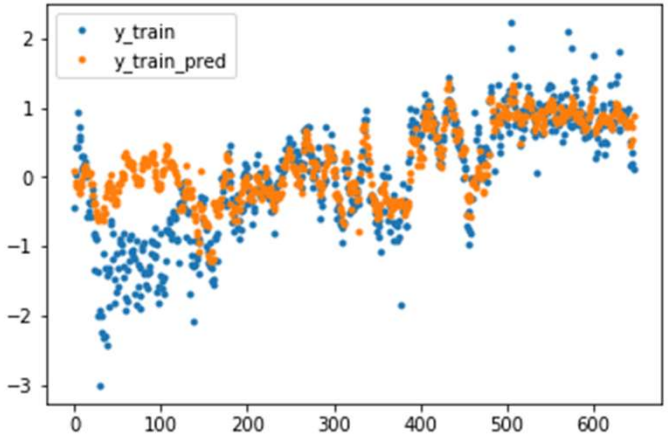
These 4 weeks have relatively high temperatures,
although they belong to S, T, T, W

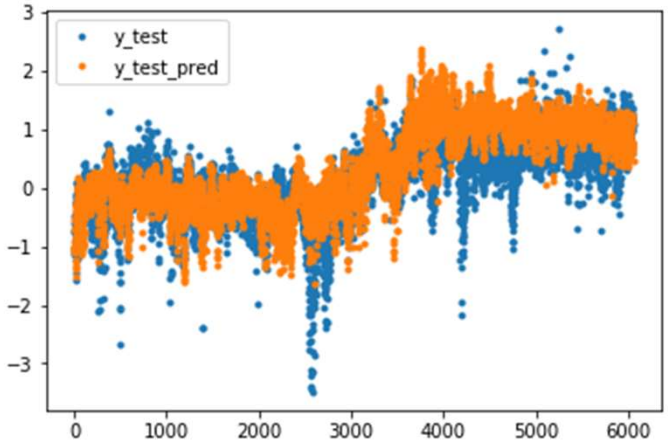
Support Vector Machine:		
EF (y_train + y_test_pred), mean, std:	**10.6**	**(13.1)**
Extrapolation for 10 months from y_train_pred:	8.7	(1.3)
Extrapolation for 10 months from y_train:	8.8	(1.4)

Units: kg per LU and per measuring period (10 months)

Prediction with Support Vector Machine (Exp.-Method 4)





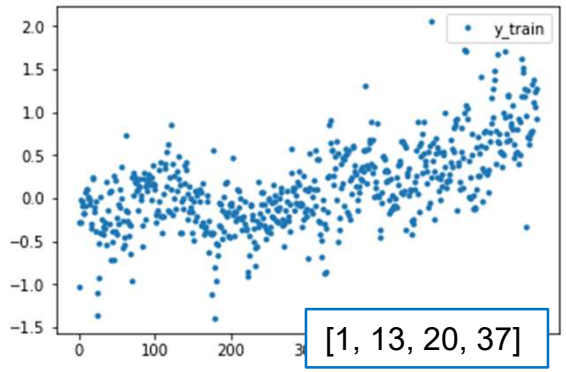


Which choice of 4 weeks out of 40

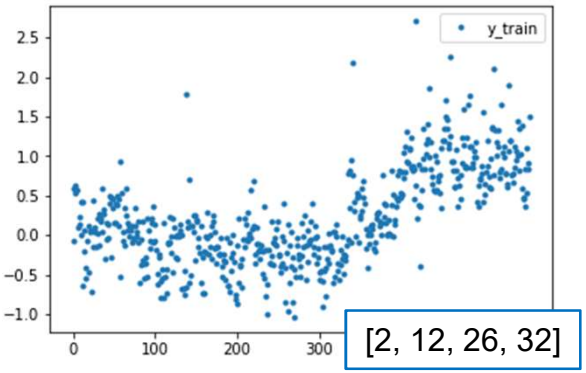
(with restriction W, S, T, T)

gives good prediction ??

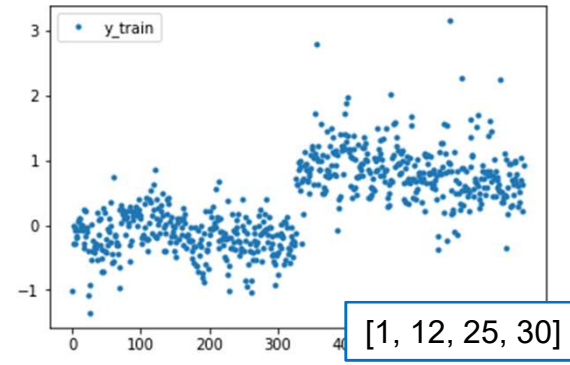
Average **EF**s



[1, 13, 20, 37]

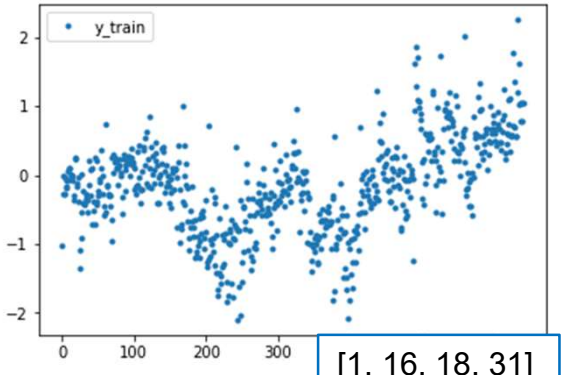


[2, 12, 26, 32]

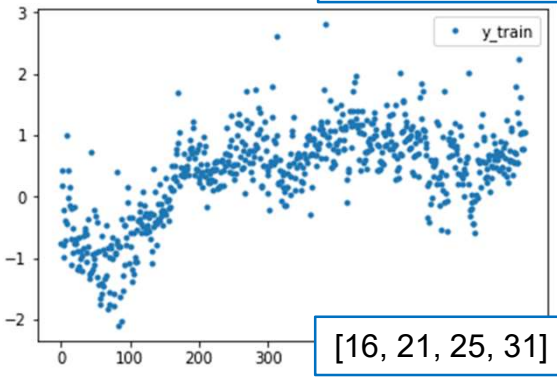


[1, 12, 25, 30]

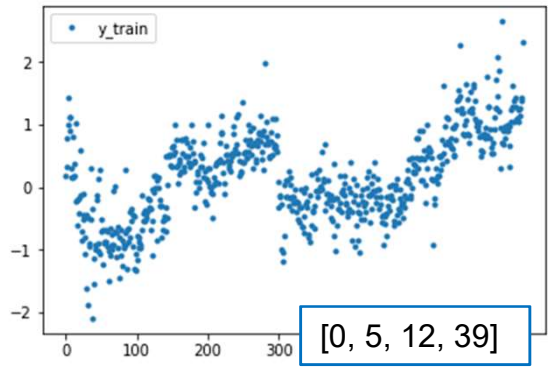
EFs below mean



[1, 16, 18, 31]

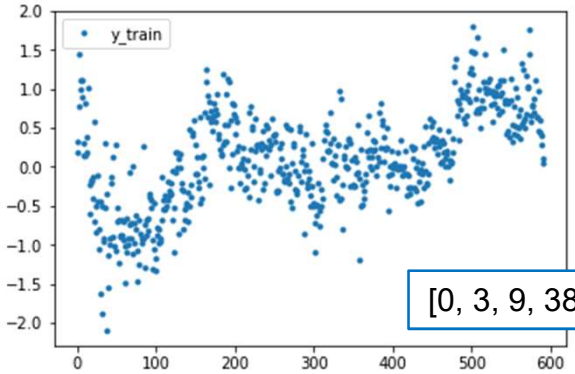


[16, 21, 25, 31]

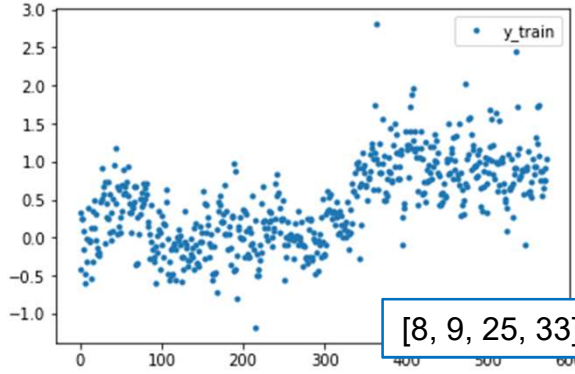


[0, 5, 12, 39]

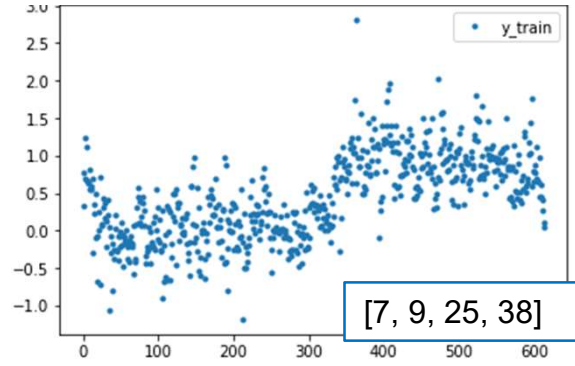
EFs above mean



[0, 3, 9, 38]

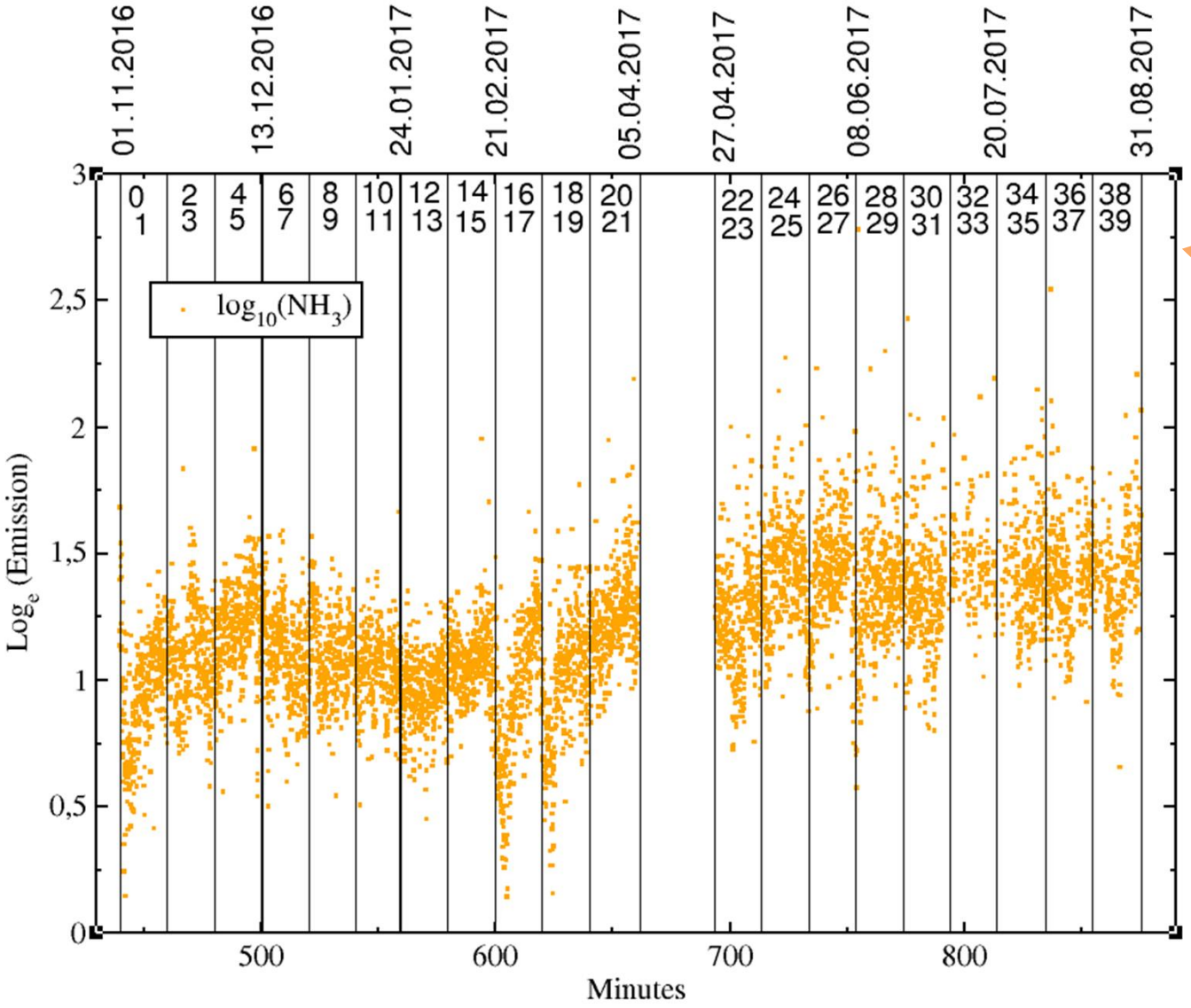


[8, 9, 25, 33]



[7, 9, 25, 38]

Dummerstorf Data

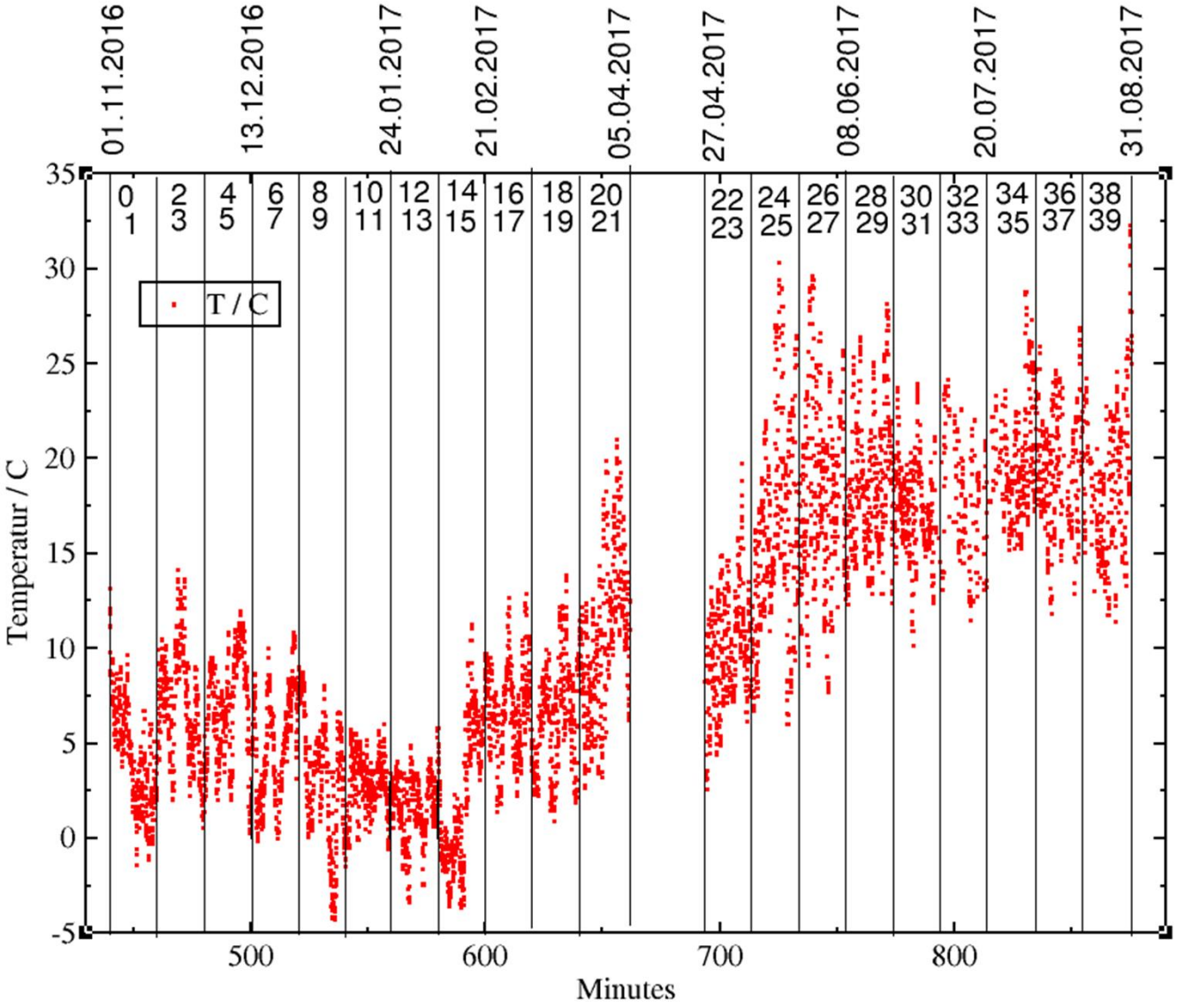


40 Intervals
of 1 weeks each

<u>Remark</u>:

On the y-axis the Logarithm of the measured emission is shown!

Dummerstorf Data



Summary and Outlook:

Is Machine Learning needed / useful in our case of emission analysis?

It seems that the methods ***polynomial regression*** and
support vector machine are a bit better than linear regression.

In our case of one barn the advantages of ML are not very convincing.

If we could analyze similar data from differenet barns, we could try if
Multi Task Learning could improve the results.

In situations of ***big*** amounts of ***data*** from different sources ML could be superior.

Machine Learning is a promising technique in many fields of science and technologie!

Combinations of classical simulations and ML can be useful in many fields, for instance Quantum Chemistry, Computational Fluid Dynamics, Proteinfolding, …

Goto Jupyter Emission Notebooks !