Lead Prediction at Flint

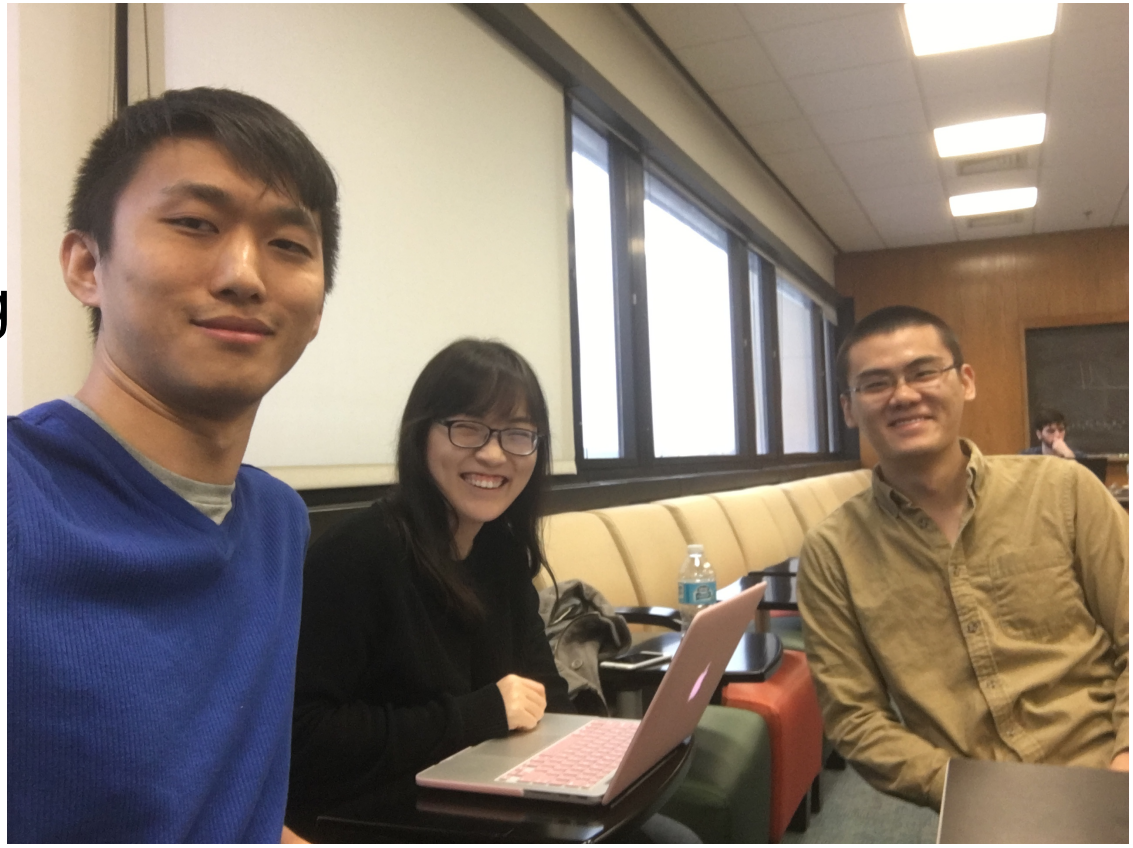
Team Pizza-Pita

Feb 26, 2017

Our Team: Pizza-Pita

Huan Chen
MS Stat

Shujian Liu
PhD MechEng



Tangxin Jin
PhD Math

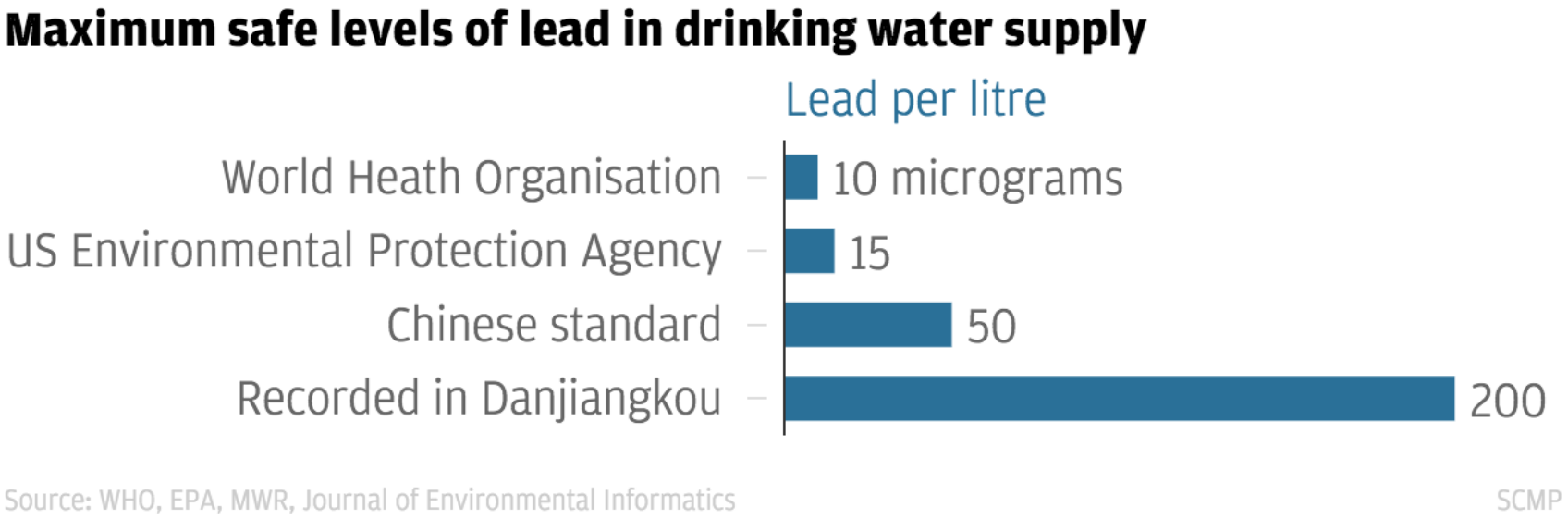
Outline

- Background
- Exploration Datasets
- Regression
  - Azure ML
  - Point to Point Modeling
  - Sequence Modeling
- New Metric
- Results
- Conclusions
- Acknowledge

Background

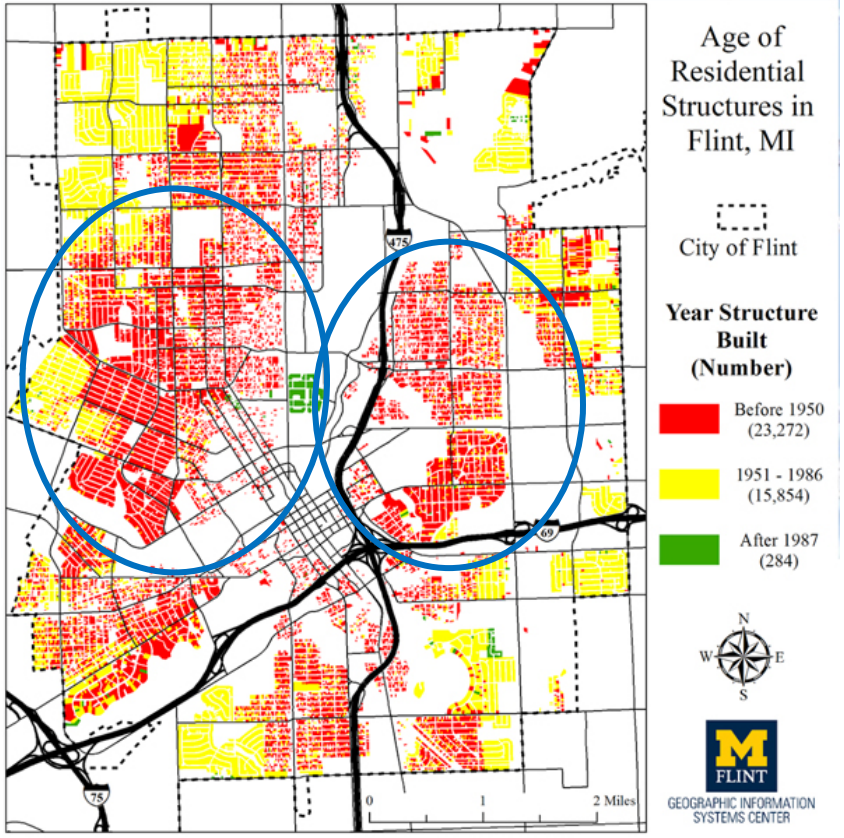
Source: http://northdallasgazette.com/2016/03/20/flint-water-crisis-move-children-babies/

Safe Levels of Lead in Water

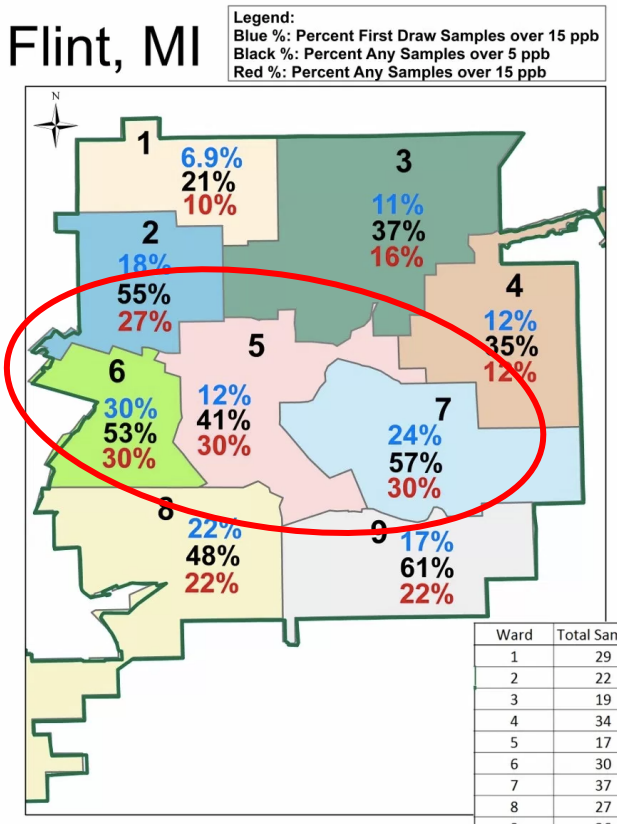


House Age and Lead Measurement

Old House



Age of Residential Structure



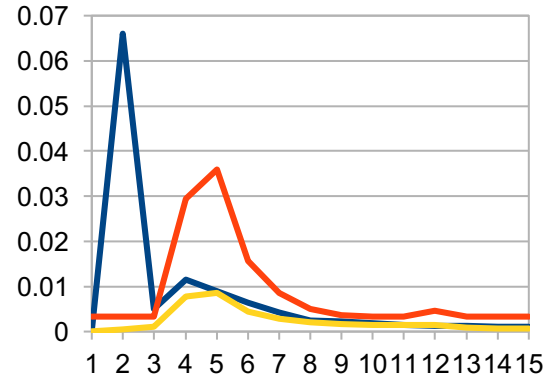
Ward	Total Samples
1	29
2	22
3	19
4	34
5	17
6	30
7	37
8	27
9	36
TOTAL	251

Lead Measurement

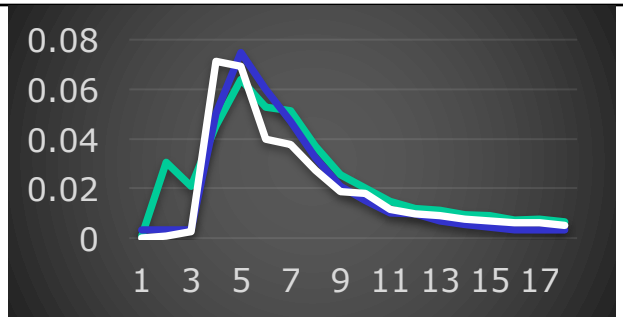
High lead

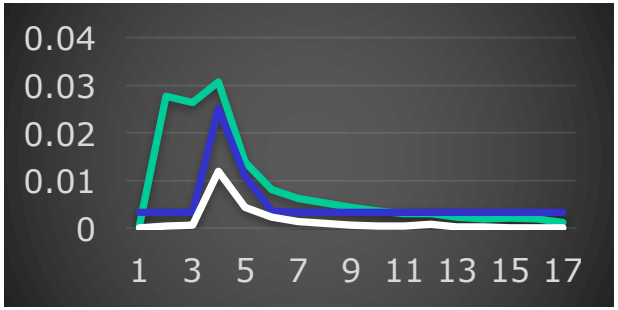
Source:
http://www.fondriest.com/news/amid-flint-water-crisis-gis-effort-maps-citys-pipes.htm
http://flintwaterstudy.org/2015/09/distribution-of-lead-results-across-flint-by-ward-and-zip-codes/

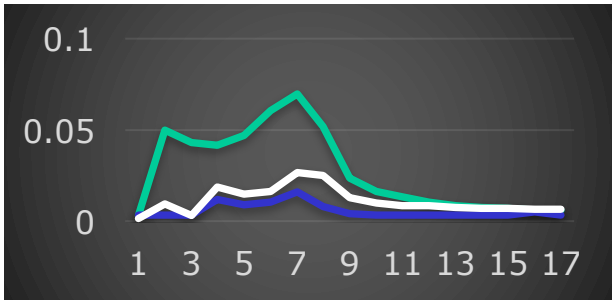
Normalized Raw Data



- Good

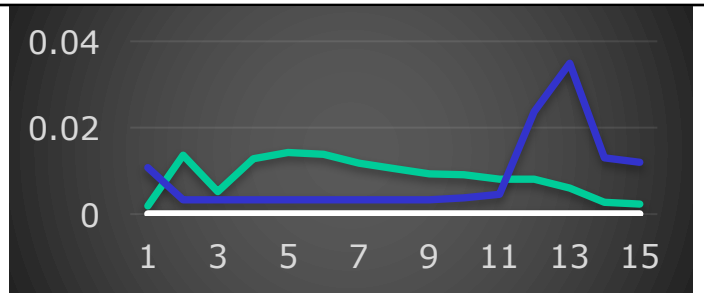


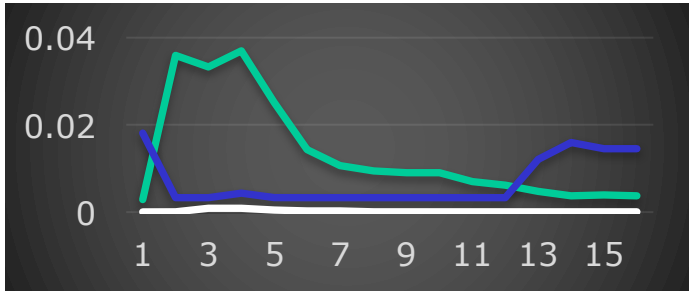


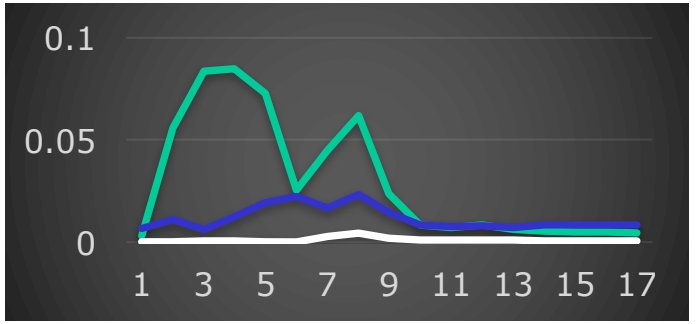


Green: Copper
Blue: Iron
White: Lead

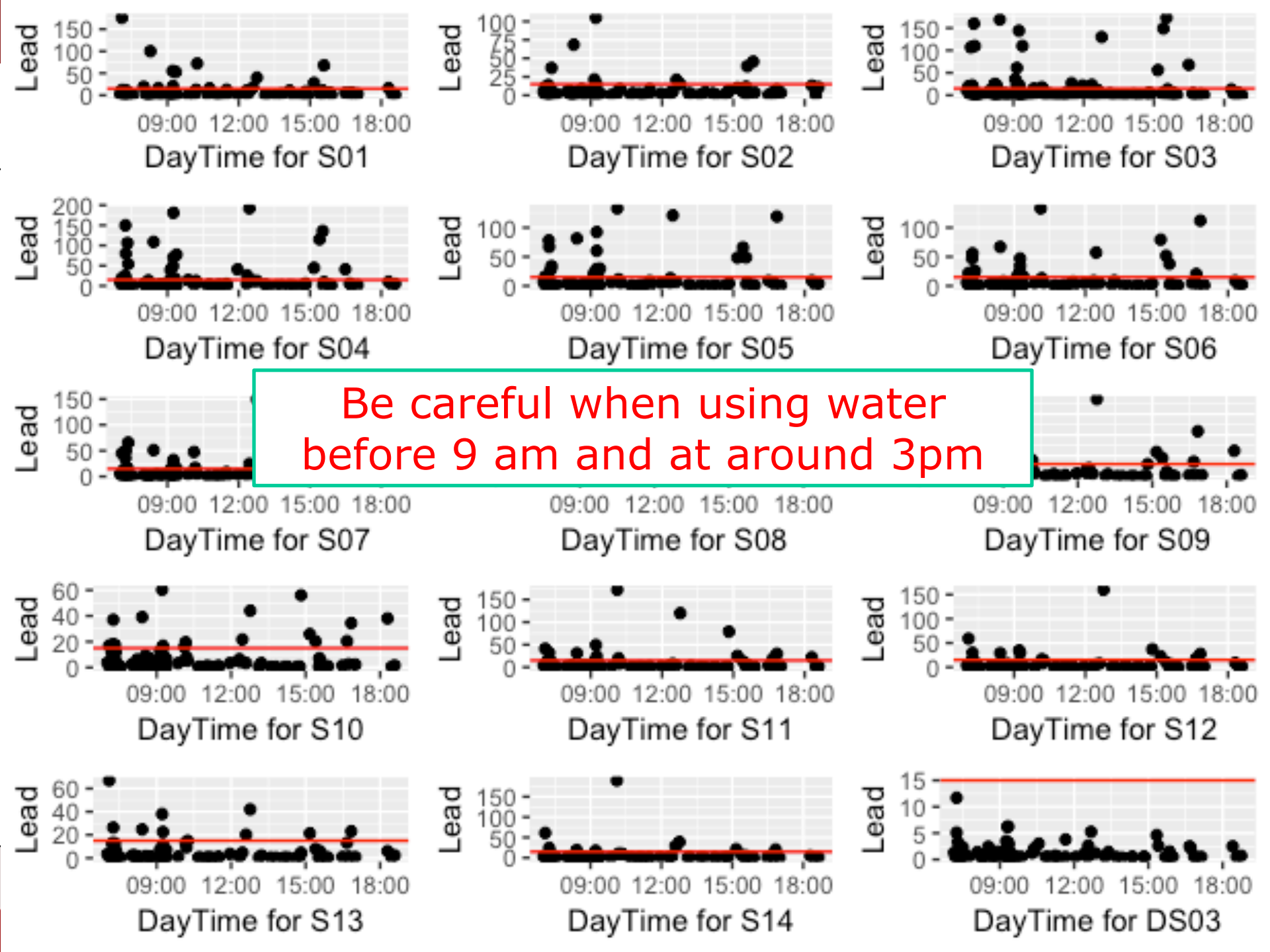
- Hum…

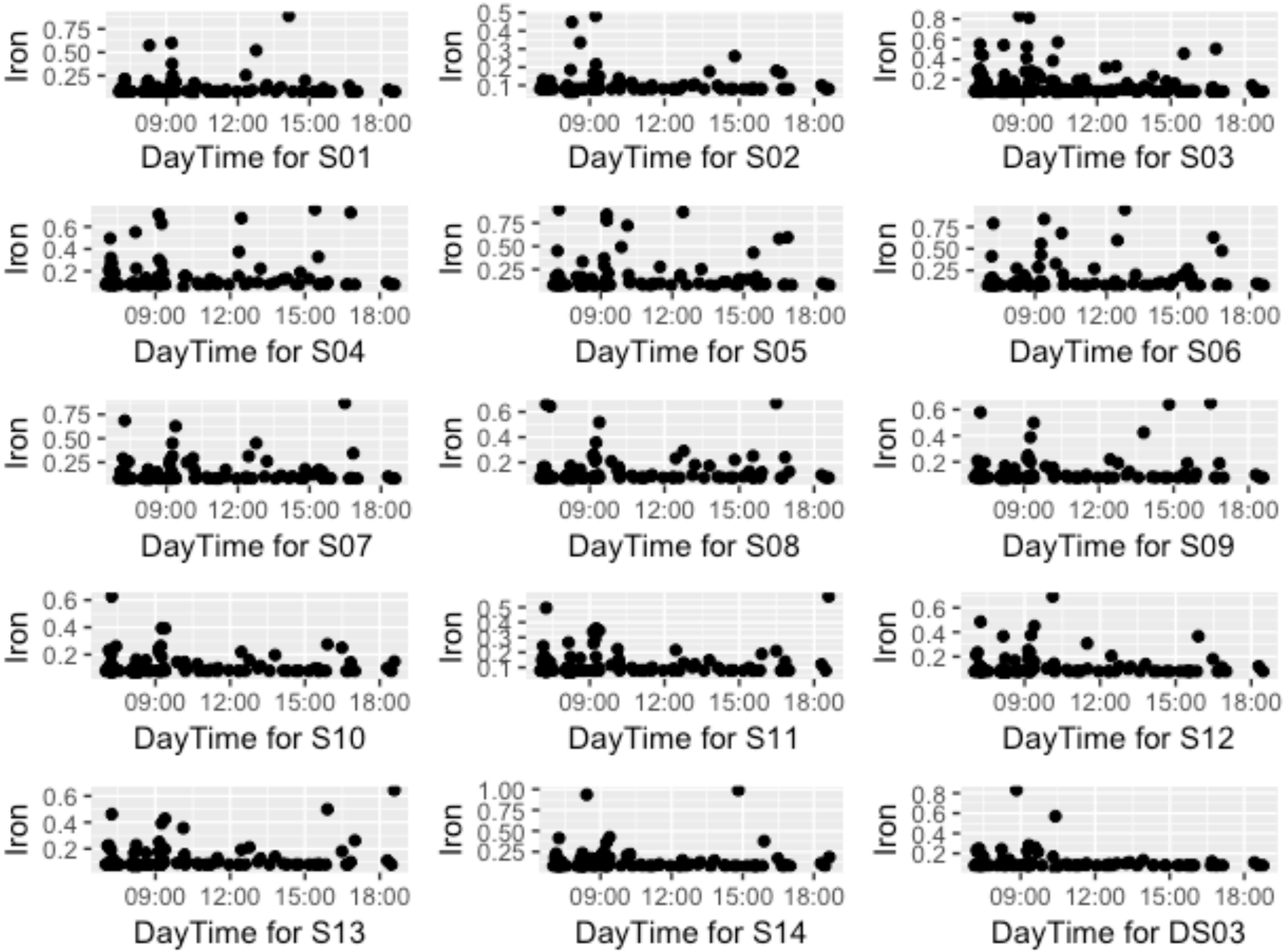






Lead


Lead
DayTime for S01
DayTime for S02
DayTime for S03
DayTime for S04
DayTime for S05
DayTime for S06
DayTime for S07
DayTime for S08
DayTime for S09
DayTime for S10
DayTime for S11
DayTime for S12
DayTime for S13
DayTime for S14
DayTime for DS03
09:00 12:00 15:00 18:00
Be careful when using water
before 9 am and at around 3pm

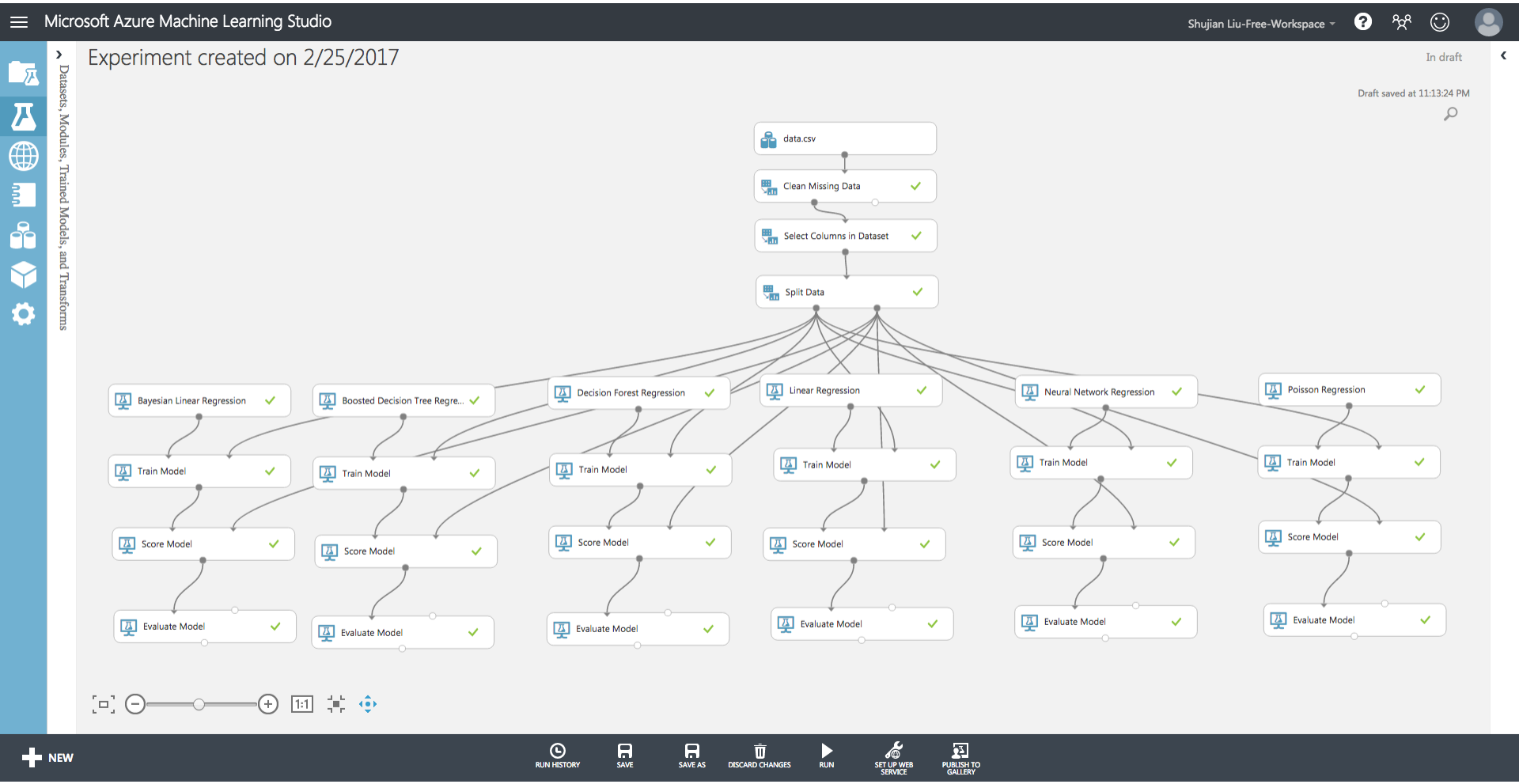

Iron
0.75
0.50
0.25
09:00 12:00 15:00 18:00
DayTime for S01
Iron
0.5
0.4
0.3
0.2
0.1
09:00 12:00 15:00 18:00
DayTime for S02
Iron
0.8
0.6
0.4
0.2
09:00 12:00 15:00 18:00
DayTime for S03
Iron
0.6
0.4
0.2
09:00 12:00 15:00 18:00
DayTime for S04
Iron
0.75
0.50
0.25
09:00 12:00 15:00 18:00
DayTime for S05
Iron
0.75
0.50
0.25
09:00 12:00 15:00 18:00
DayTime for S06
Iron
0.75
0.50
0.25
09:00 12:00 15:00 18:00
DayTime for S07
Iron
0.6
0.4
0.2
09:00 12:00 15:00 18:00
DayTime for S08
Iron
0.6
0.4
0.2
09:00 12:00 15:00 18:00
DayTime for S09
Iron
0.6
0.4
0.2
09:00 12:00 15:00 18:00
DayTime for S10
Iron
0.5
0.1
09:00 12:00 15:00 18:00
DayTime for S11
Iron
0.6
0.4
0.2
09:00 12:00 15:00 18:00
DayTime for S12
Iron
0.6
0.4
0.2
09:00 12:00 15:00 18:00
DayTime for S13
Iron
1.00
0.75
0.50
0.25
09:00 12:00 15:00 18:00
DayTime for S14
Iron
0.8
0.6
0.4
0.2
09:00 12:00 15:00 18:00
DayTime for DS03

Data Manipulation

- Microsoft Azure Machine Learning Studio:
  - 4 variables
  - 75% of data for training
- Point-point Modeling (scikit-learn):
  - 2 variables: training-testing split 1000/443
  - 4 variables: training-testing split 900/368
- Sequence Modeling (Keras on TF):
  - Only used data from 15 samples (67 sequences)
  - 40 sequences data for training
  - 10% from training for cross-validation
  - 27 sequences data for testing

Azure ML

Microsoft Azure



Results from Azure

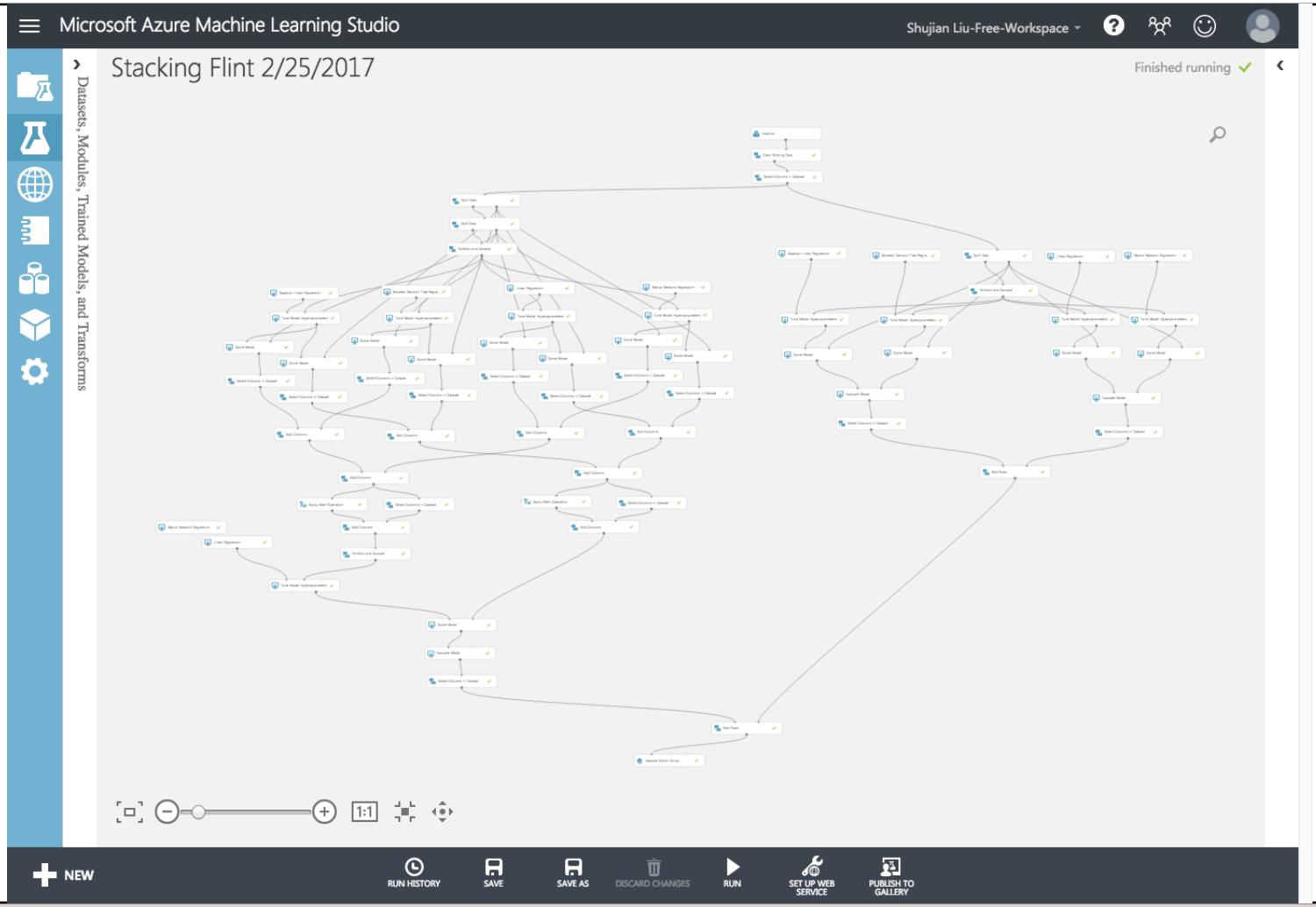
Microsoft Azure

Models	Mean Absolute Error	Root Mean Squared Error
Bayesian Linear Regression	21.263156	155.605892
Boosted Decision Tree	18.345794	157.411814
Decision Forest Regression	16.235807	157.512419
Linear Regression	21.581666	155.500647
Neural Network	22.395735	158.531654
Poisson Regression	23.83821	158.520784

Machine Learning in ONE HOUR

Ensemble Regressors using Stacking

Microsoft Azure



Results from Stacking

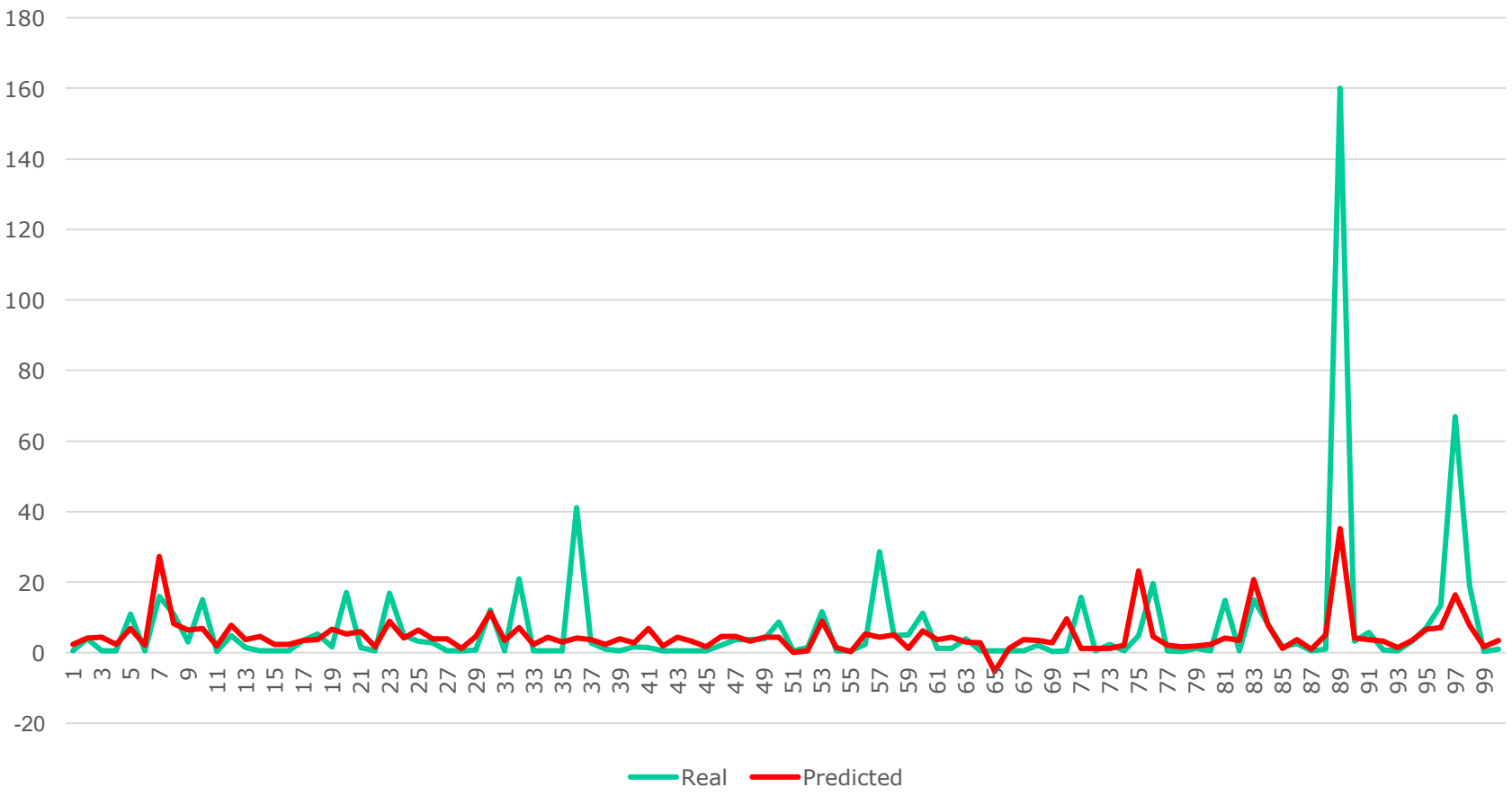
- Stacking of four basic regression algorithms:
  - Bayesian Linear Regression
  - Boosted Decision Tree Regression
  - Linear Regression
  - Neural Network Regression
- Metrics:

Mean Absolute Error	5.118008
Root Mean Squared Error	13.088811

- Compare to Decision Forest Regression with MAE of 16.2

Reference: https://gallery.cortanaintelligence.com/Experiment/Building-Ensemble-of-Classifiers-using-Stacking-2
Breiman, Leo. "Stacked regressions." *Machine learning* 24.1 (1996): 49-64.

Results from Stacking



Point to Point Modeling (scikit-learn)

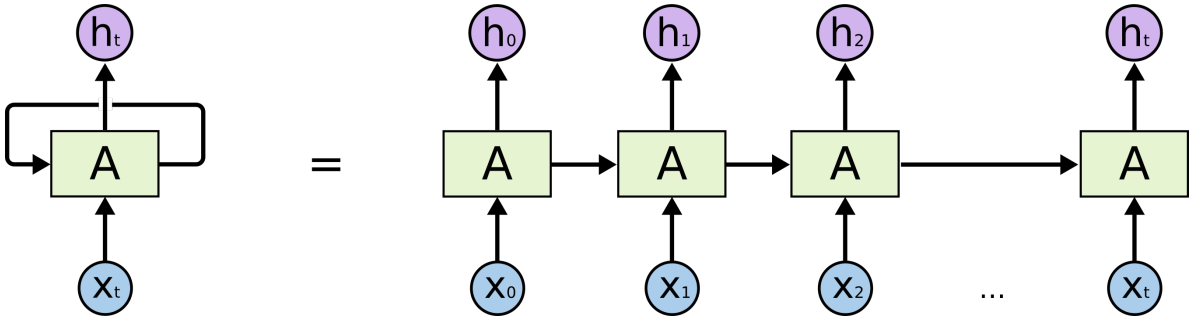
Training with 2 variables
(copper and iron)

Model	Linear regression	Ridge	Lasso	SVM (rbf)	Random Forest (5 trees)
MAE	14.21	14.21	16.90	15.86	20.70

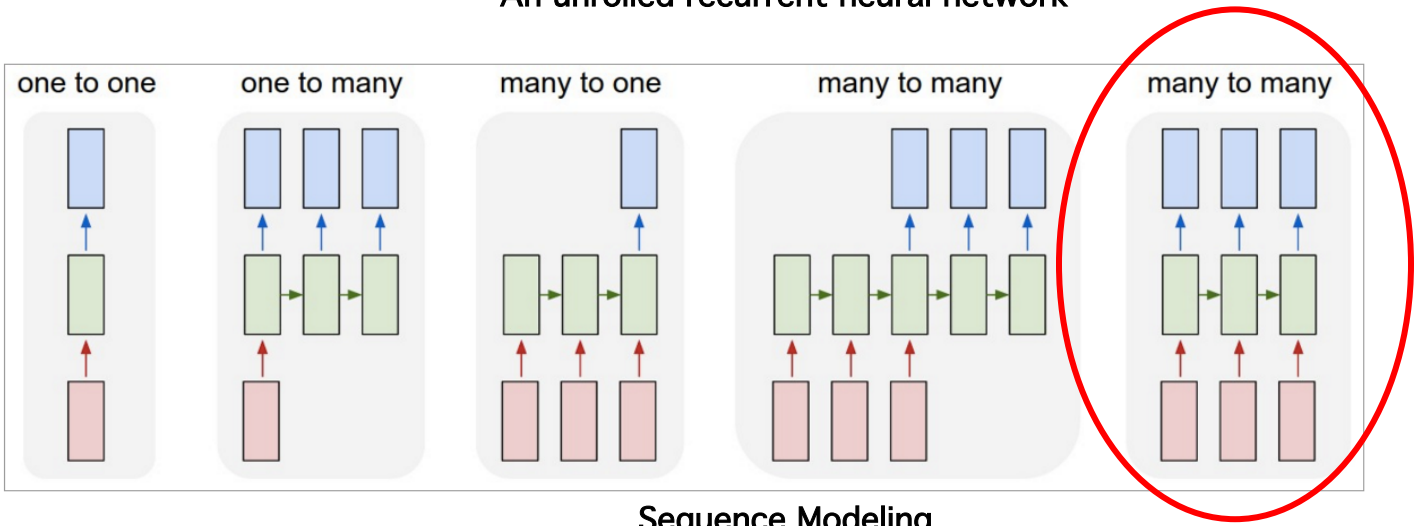
Training with 4 variables
(copper, iron, Cl2, pH)

Model	Linear regression	SVM (rbf)	Random Forest (5 trees)
MAE	19.21	16.78	18.94

Recurrent Neural Network / Long Short Term Memory

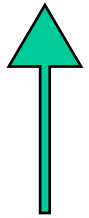


An unrolled recurrent neural network



Sequence Modeling

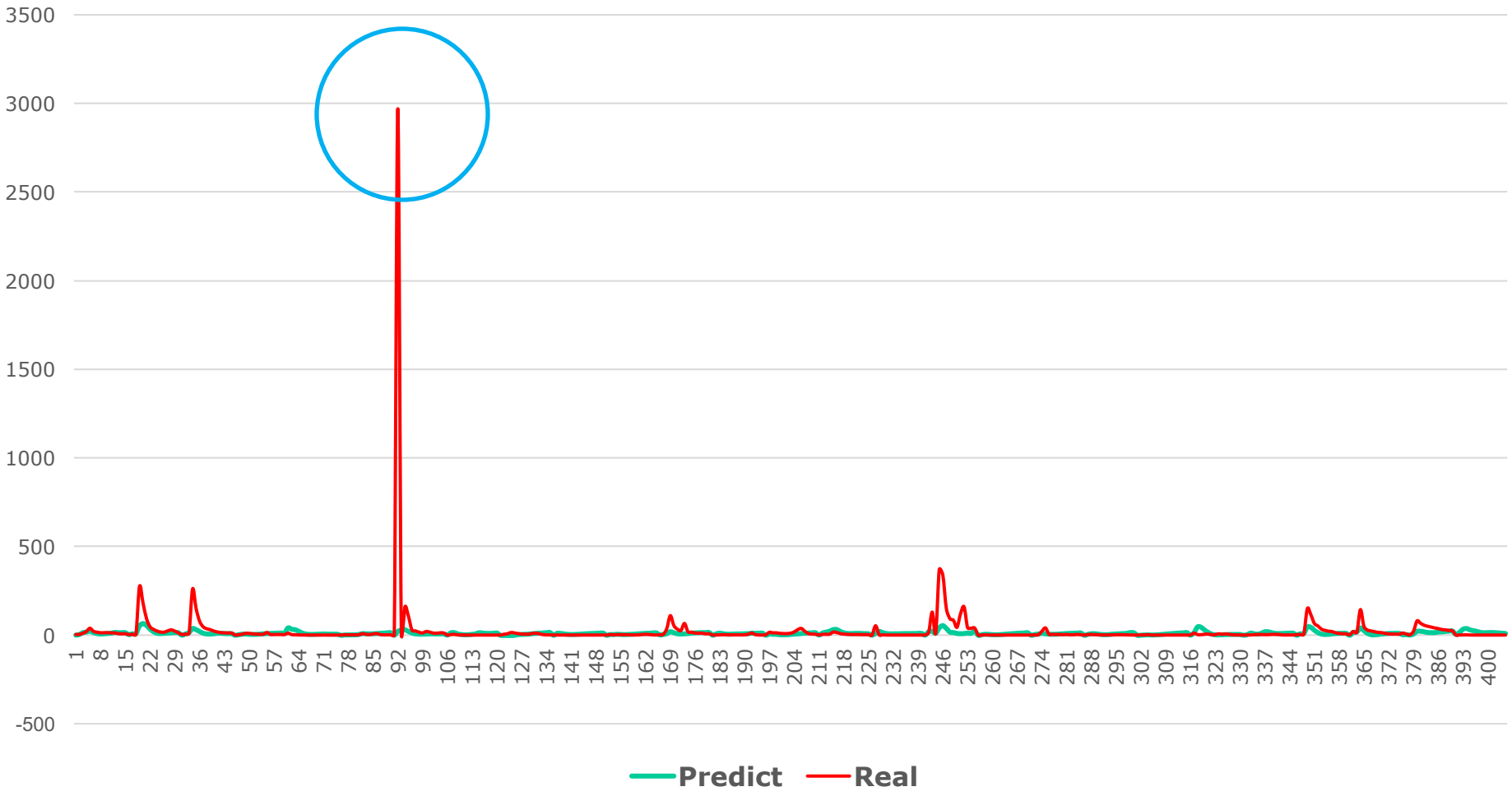
Lead



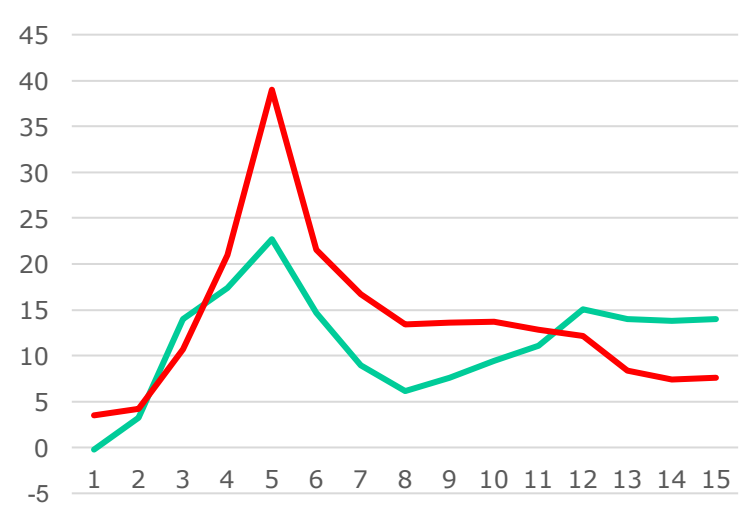
Copper & Iron

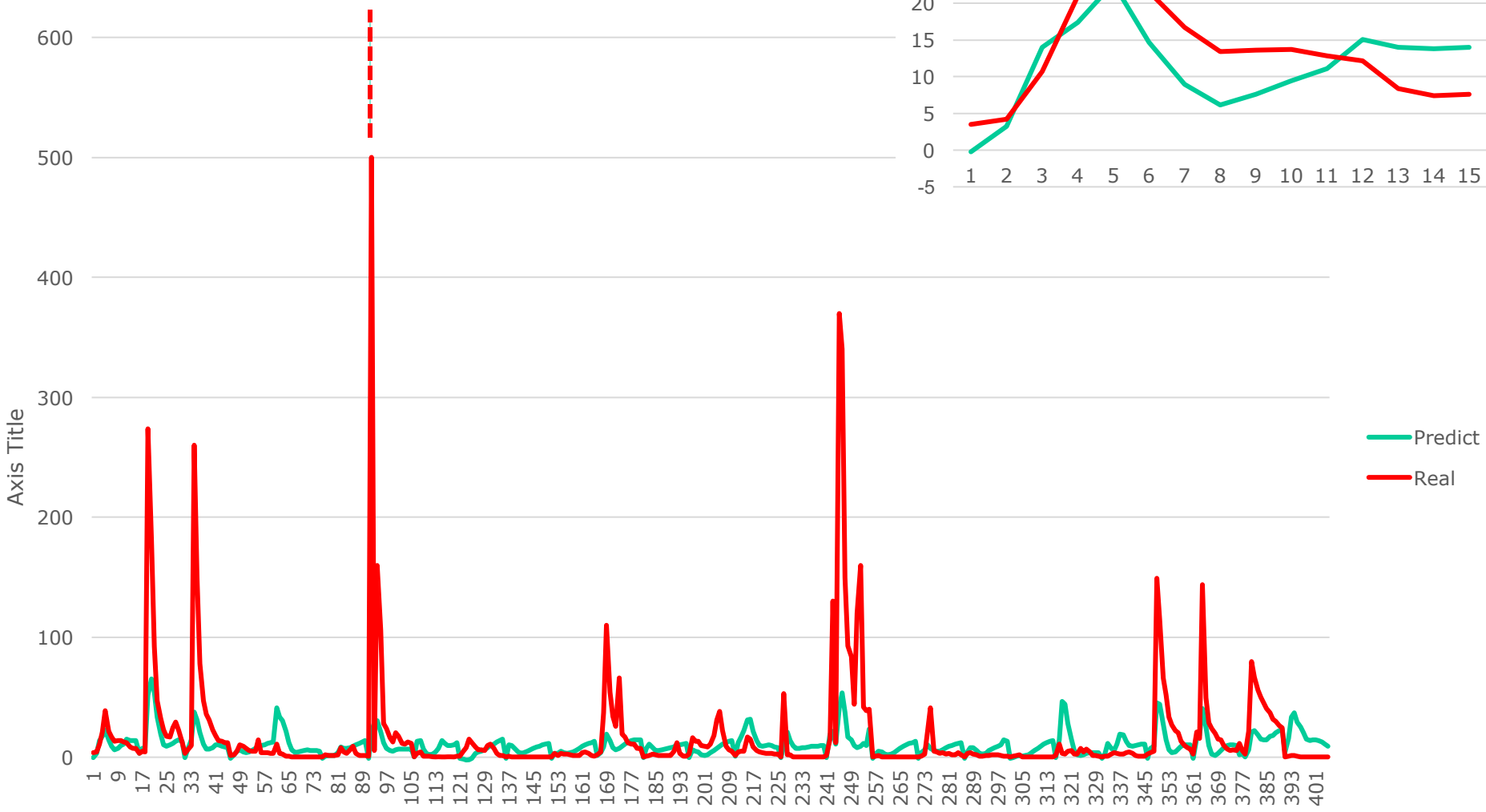
Source:
http://colah.github.io/posts/2015-08-Understanding-LSTMs/
http://karpathy.github.io/2015/05/21/rnn-effectiveness/

Lead Prediction by RNN/LSTM



RNN/LSTM


45
40
35
30
25
20
15
10
5
0
-5
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15


600
500
400
300
200
100
0
-100
Axis Title
1 9 17 25 33 41 49 57 65 73 81 89 97 105 113 121 129 137 145 153 161 169 177 185 193 201 209 217 225 233 241 249 257 265 273 281 289 297 305 313 321 329 337 345 353 361 369 377 385 393 401
Predict
Real

New Metric

New Metric:
Instead of MAE/RMSE loss, we tried rate of successfully predicted safety level (15 ug/L).
~ in classification

New Metric – 15 ug/L

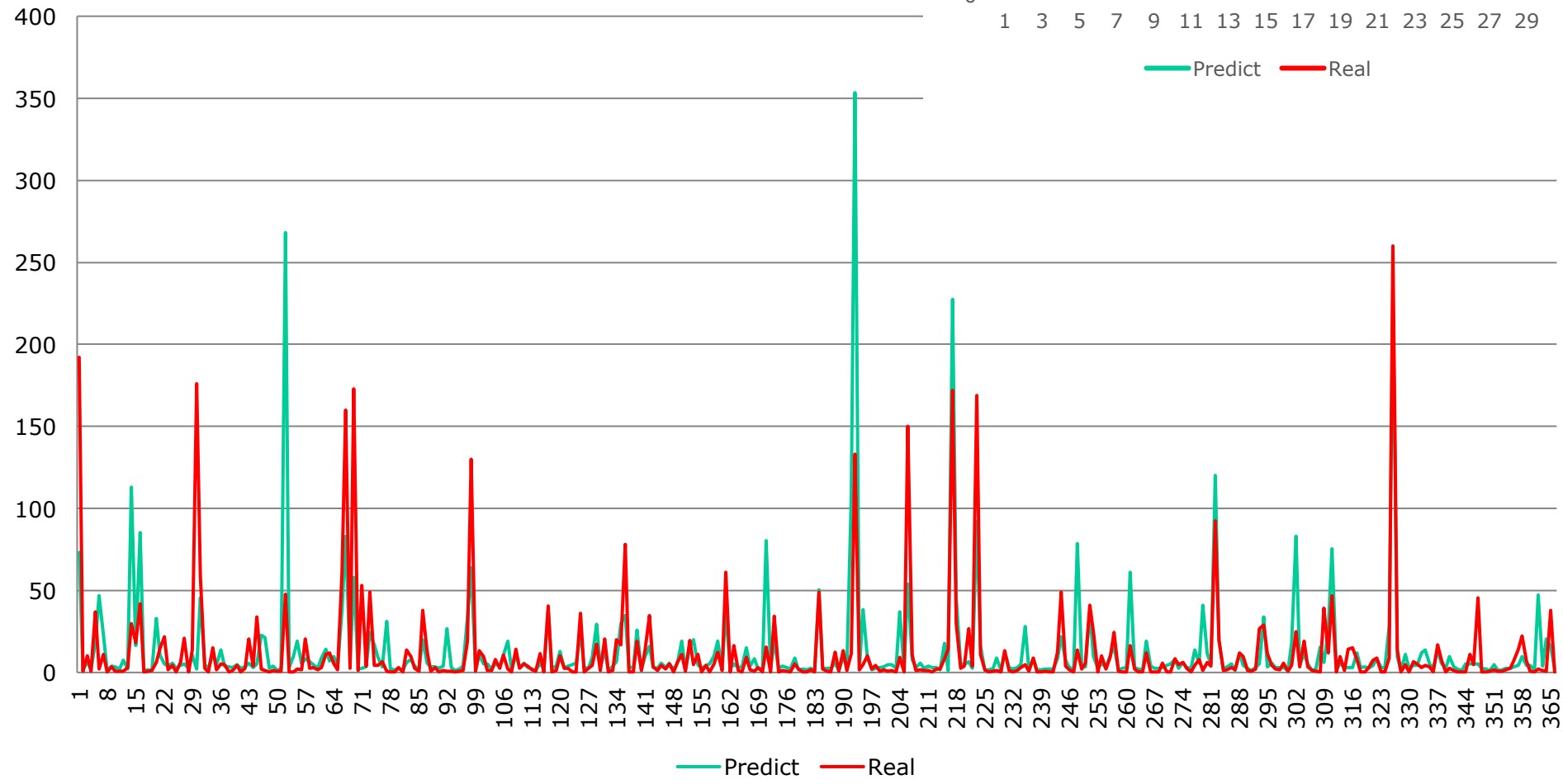
Training with 2 Variables (Copper and Iron)

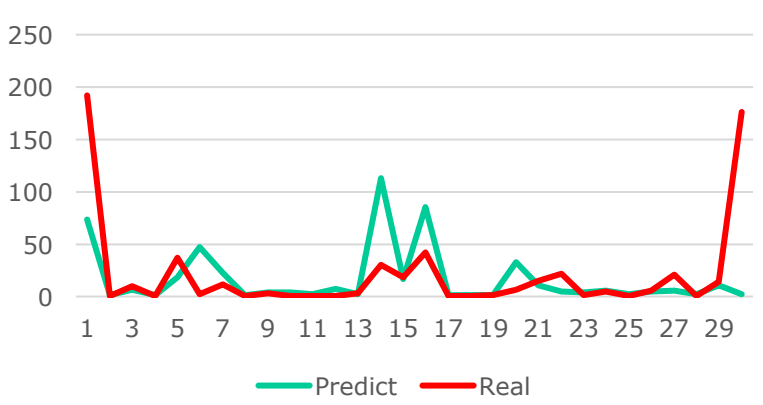
Model	Linear regression	SVM (rbf)	Random Forest (5 trees)	RNN/LSTM (10 epochs)
Succ. Rate	0.72	0.80	0.84	0.82

Training with 4 Variables

Model	Linear regression	SVM (rbf)	Random Forest (10 trees)	Stacking Azure ML
Succ. Rate	0.69	0.78	0.87	0.89 !

Random Forest





Conclusion

- No strong linear correlation between copper/iron with lead is found.
- Be careful when using water before 9 am and at around 3pm.
- Training with 4 variables (copper, iron, Cl2 and pH) tends to behave better than with only copper and iron.
- Stacking model has the best performance. Random Forest algorithm comes second.
- RNN/LSTM has great potential but large dataset is needed.

Acknowledge

- Grid Club
  - Great food!
- Mentors
- Sponsors
- Microsoft Azure