Machine Learning in Air Quality: Overview of applications and a case study on the SmartAQ forecasting system

Dr. Ioannis D. Apostolopoulos
Postdoctoral Researcher, FORTH/ICE-HT

FORTH

C·STACC
Center for the Study of Air Quality & Climate Change

04.01.24

Artificial Intelligence and Machine Learning

ARTIFICIAL INTELLIGENCE

A program that can sense, reason, act, and adapt.

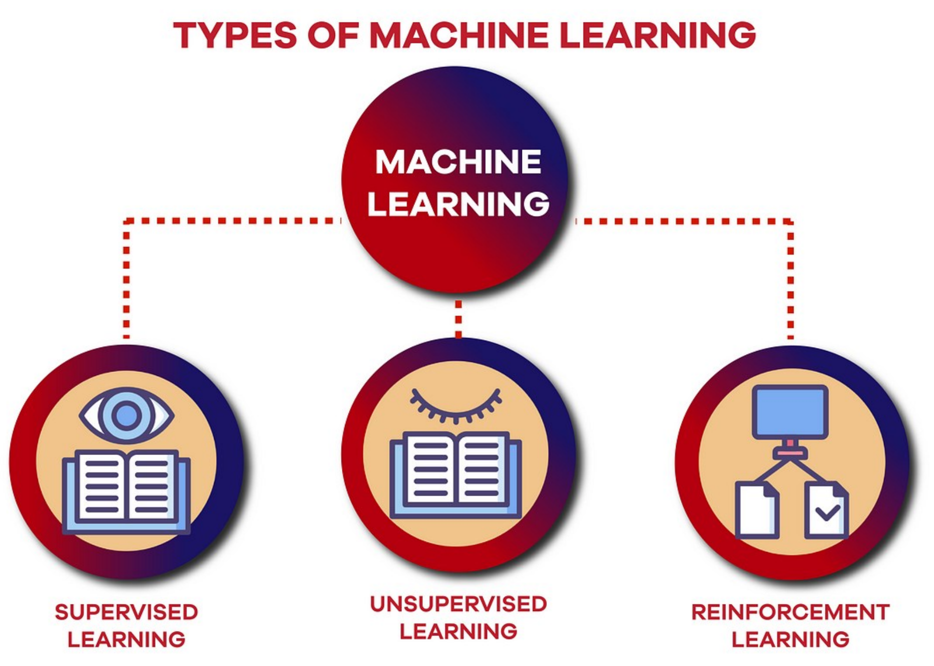
MACHINE LEARNING

Algorithms whose performance improve as they are exposed to more data over time.

DEEP LEARNING

Subset of machine learning in which multilayered neural networks learn from vast amount of data.

Source: www.globaltechcouncil.org



TYPES OF MACHINE LEARNING

MACHINE LEARNING

SUPERVISED LEARNING

UNSUPERVISED LEARNING

REINFORCEMENT LEARNING

Basic principles of some algorithms

Naïve Bayes
Classify data points based on the probability of belonging to a particular class.

Linear Regression – Fit a line

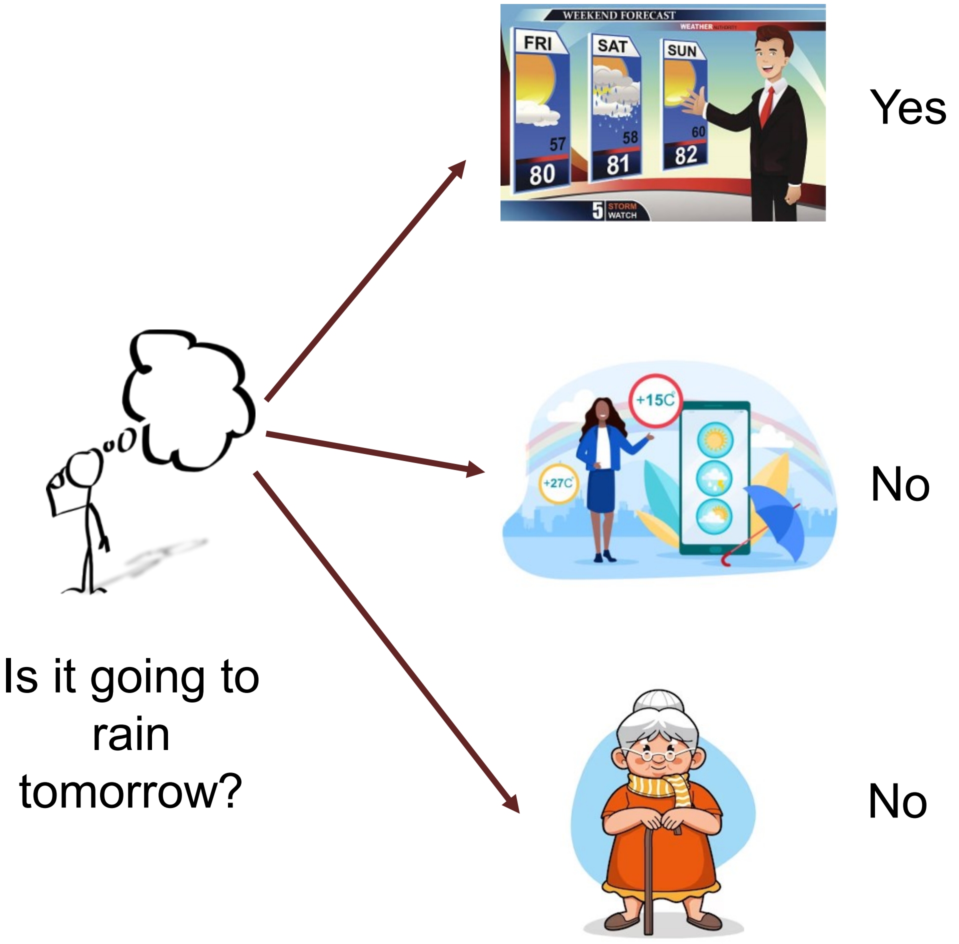
Support Vector Machines (SVM)
Find the hyperplane that best separates different classes in a high-dimensional space.

K-Nearest Neighbors (K-NN)
Classify a data point based on the majority class of its k-nearest neighbors.

Decision Tree
Create a tree-like model of decisions to represent possible outcomes and their consequences.

Neural Network
Neurons and Layers. Weights the input features and updates the weights using back-propagation and gradients

What is Ensemble Learning



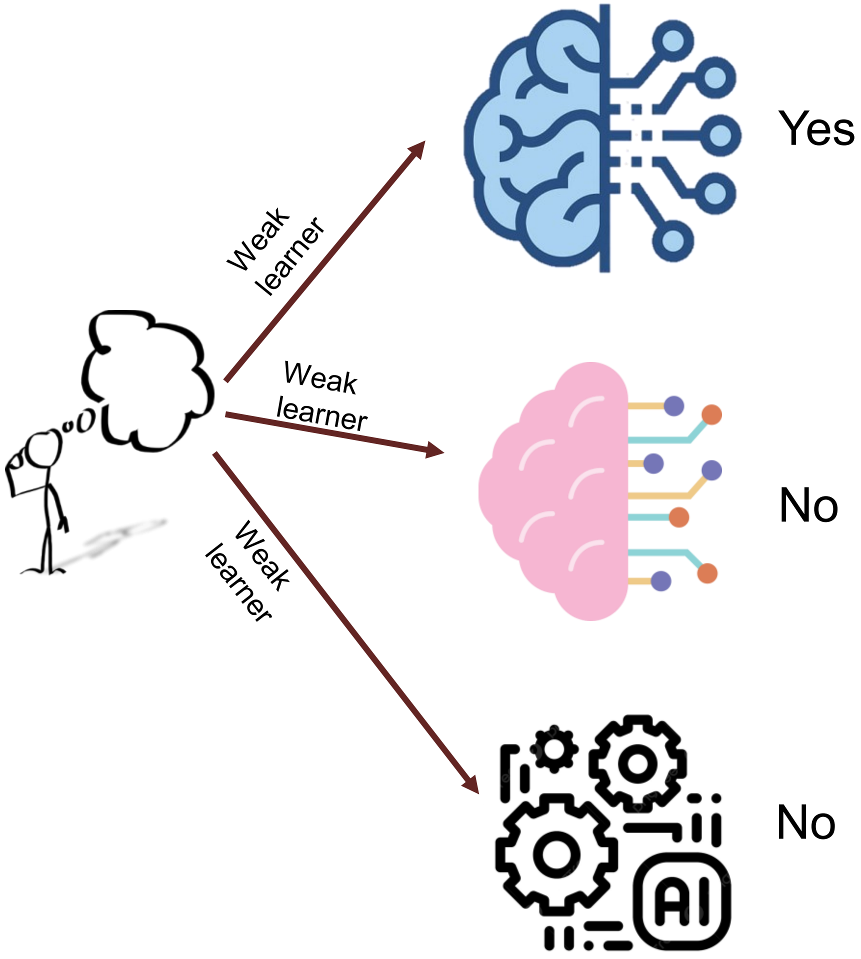
Is it going to rain tomorrow?

Yes

No

No

The majority votes for “No rain”



Weak learner

Weak learner

Weak learner

Yes

No

No

The majority votes for “No rain”

Applications in Air Quality

AQ Sensors

- Calibration
- Analysis
- Anomaly detection
- Wireless Sensor Networks (WSN)
- Security

Measurements

- Calibration
- Satellite data processing

Modelling

- Domain adaption
- Data assimilation
- Emission prediction
- Prediction improvement

Data Analytics

- Clustering
- Seasonality detection
- Outlier detection
- Discover hidden patterns

Forecasting

- Using WSNs
- Using historical measurements

Summary

- Machine Learning is **part** of Artificial Intelligence and it is not something new
- **Supervised** and **unsupervised** learning
- Multiple algorithms, multiple principles
- Ensemble Learning with XGBoost
- Various applications in AQ

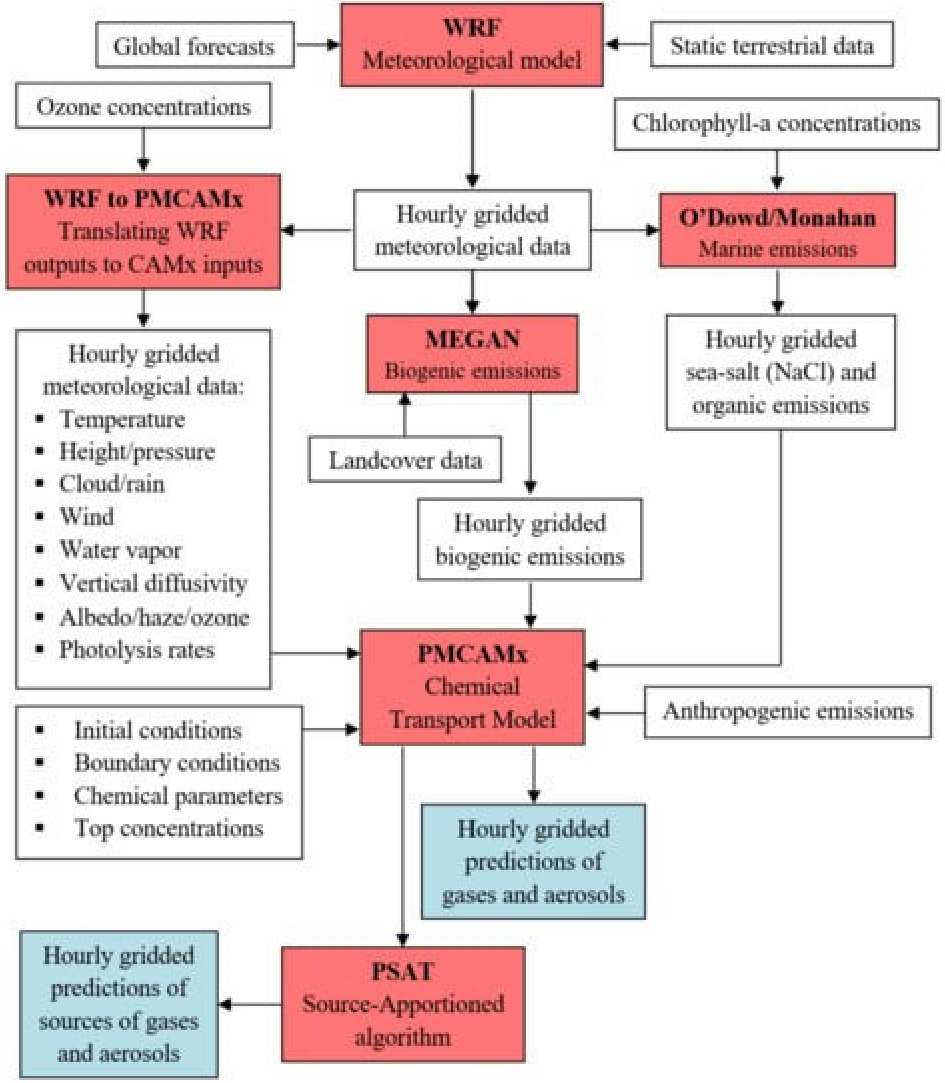
Case study

Machine Learning-enhanced Smart Air Quality forecasting system for improved estimations of gas-phase pollutants in an urban area during summer months

SmartAQ introduction

- Past works of Dr. Valia Siouti
- Collaborative efforts and outcomes from:
  - Prof. Ioannis Kioutsoukis
  - Dr. Ksakoutsi Skyllakou
  - Dr. David Patoulias
- Supervision from Prof. Spyros Pandis

- Combines meteorological and chemical transport models
- Weather Research and Forecasting (WRF) and Particulate Matter Comprehensive Air quality Model with extensions (PMCAMx)
- Provides three-day forecast of the concentration of gas-phase air pollutants



Global forecasts

WRF
Meteorological model

Static terrestrial data

Ozone concentrations

Chlorophyll-a concentrations

WRF to PMCAMx
Translating WRF outputs to CAMx inputs

Hourly gridded meteorological data

O'Dowd/Monahan
Marine emissions

Hourly gridded meteorological data:
- Temperature
- Height/pressure
- Cloud/rain
- Wind
- Water vapor
- Vertical diffusivity
- Albedo/haze/ozone
- Photolysis rates

MEGAN
Biogenic emissions

Hourly gridded sea-salt (NaCl) and organic emissions

Landcover data

Hourly gridded biogenic emissions

PMCAMx
Chemical Transport Model

Anthropogenic emissions

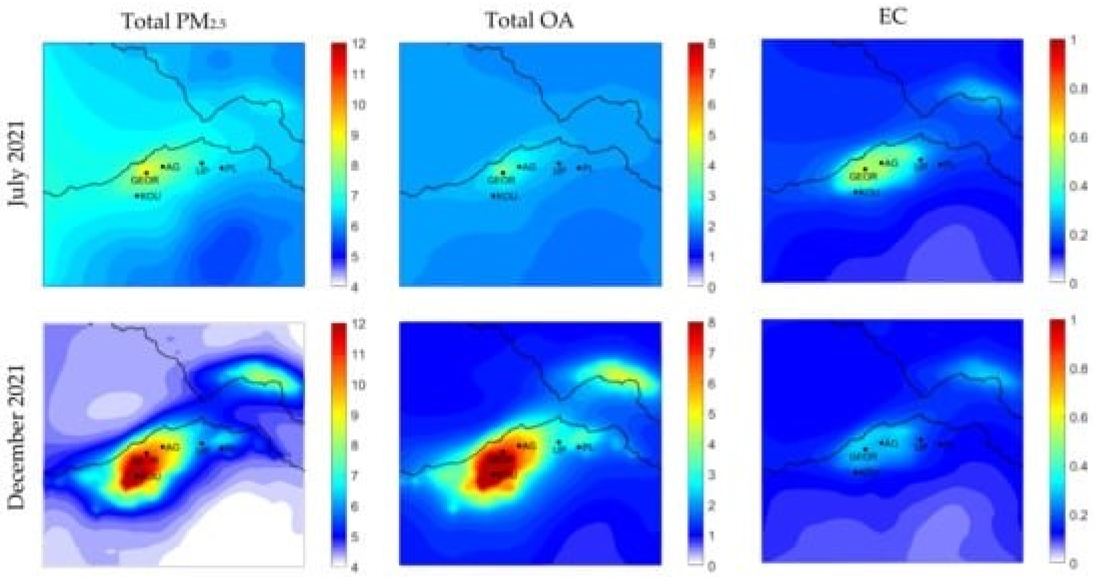
- Initial conditions
- Boundary conditions
- Chemical parameters
- Top concentrations

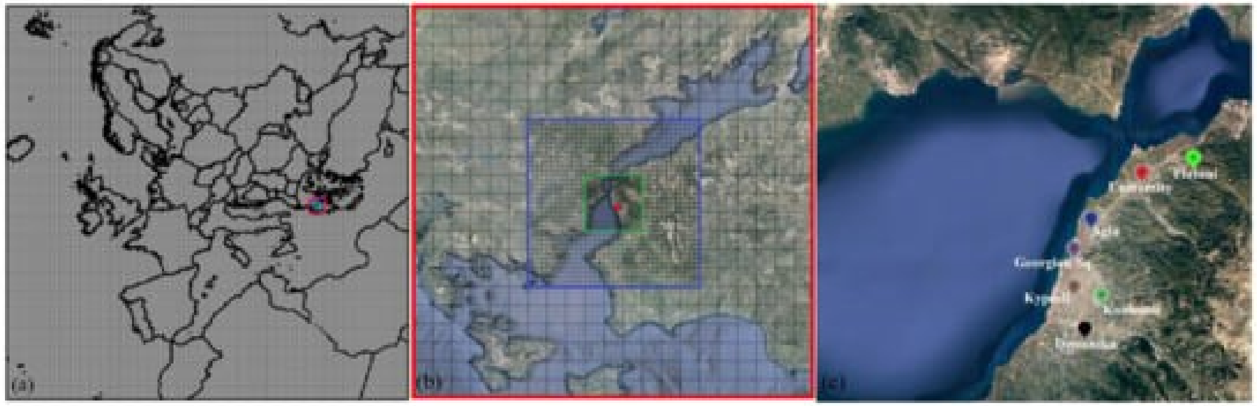
Hourly gridded predictions of gases and aerosols

Hourly gridded predictions of sources of gases and aerosols

PSAT
Source-Apportioned algorithm

SmartAQ outputs

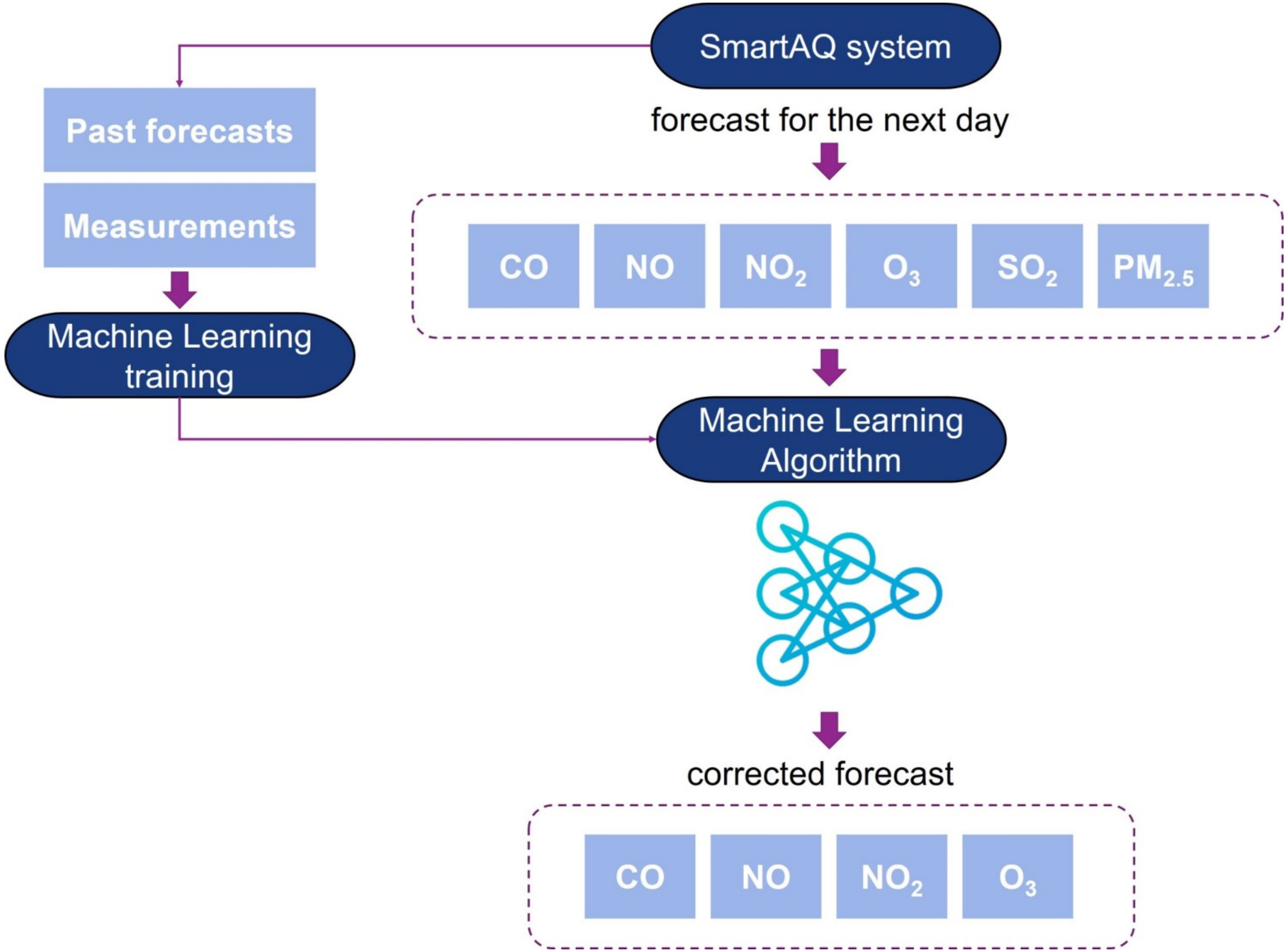




Siouti E, Skyllakou K, Kioutsioukis I, Patoulias D, Fouskas G, Pandis SN. Development and Application of the SmartAQ High-Resolution Air Quality and Source Apportionment Forecasting System for European Urban Areas. Atmosphere. 2022; 13(10):1693

- Operates in real time
- Provides three-day forecast of the concentration of gas-phase air pollutants
- Provides the complete aerosol size/composition distribution
- source contributions for all primary and secondary pollutants
- Area of 36x36 km
- Resolution of 1x1 km^2

The SmartAQ system and the proposed ML integration



SmartAQ system

Past forecasts

Measurements

Machine Learning training

forecast for the next day

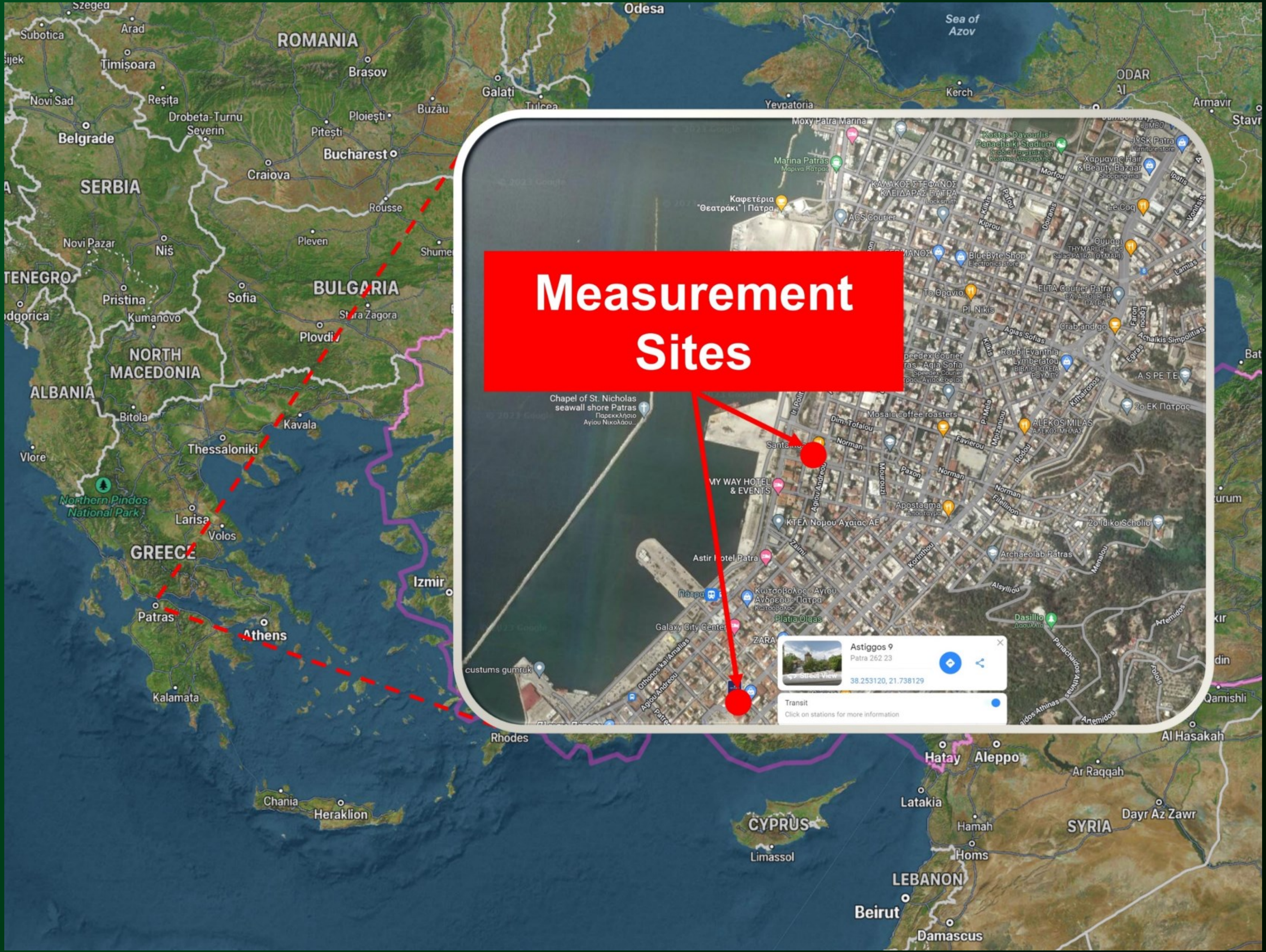
CO NO NO_2 O_3 SO_2 $PM_{2.5}$

Machine Learning Algorithm

corrected forecast

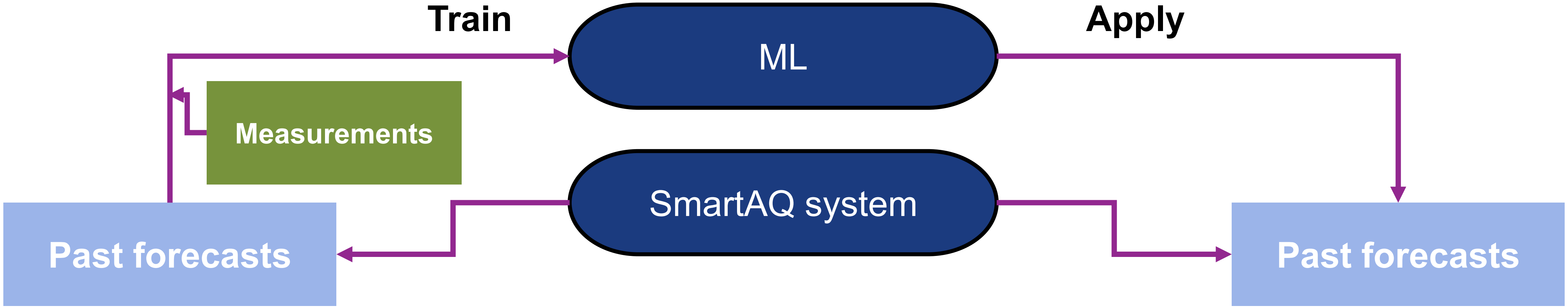
CO NO NO_2 O_3

Sites



Measurement Sites

Experiment overview



Train

ML

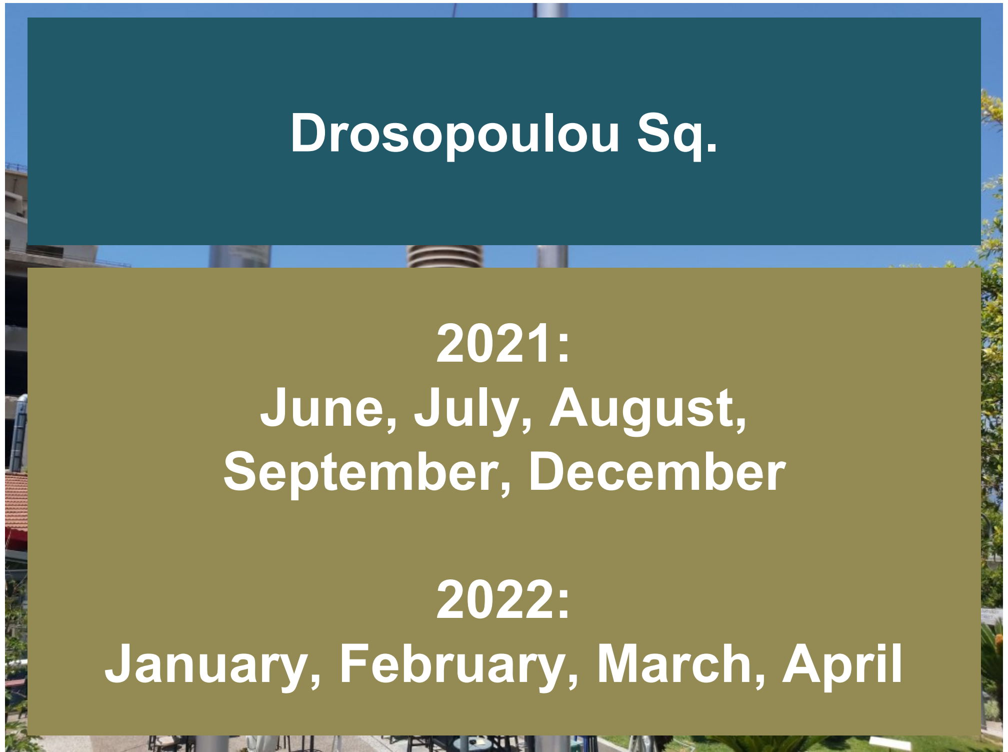
Apply

Measurements

SmartAQ system

Past forecasts

Past forecasts



Drosopoulou Sq.

2021:
June, July, August,
September, December

2022:
January, February, March, April

Georgiou Sq.

2023:
May, June, July,
August, September, October

SmartAQ operation – outputs

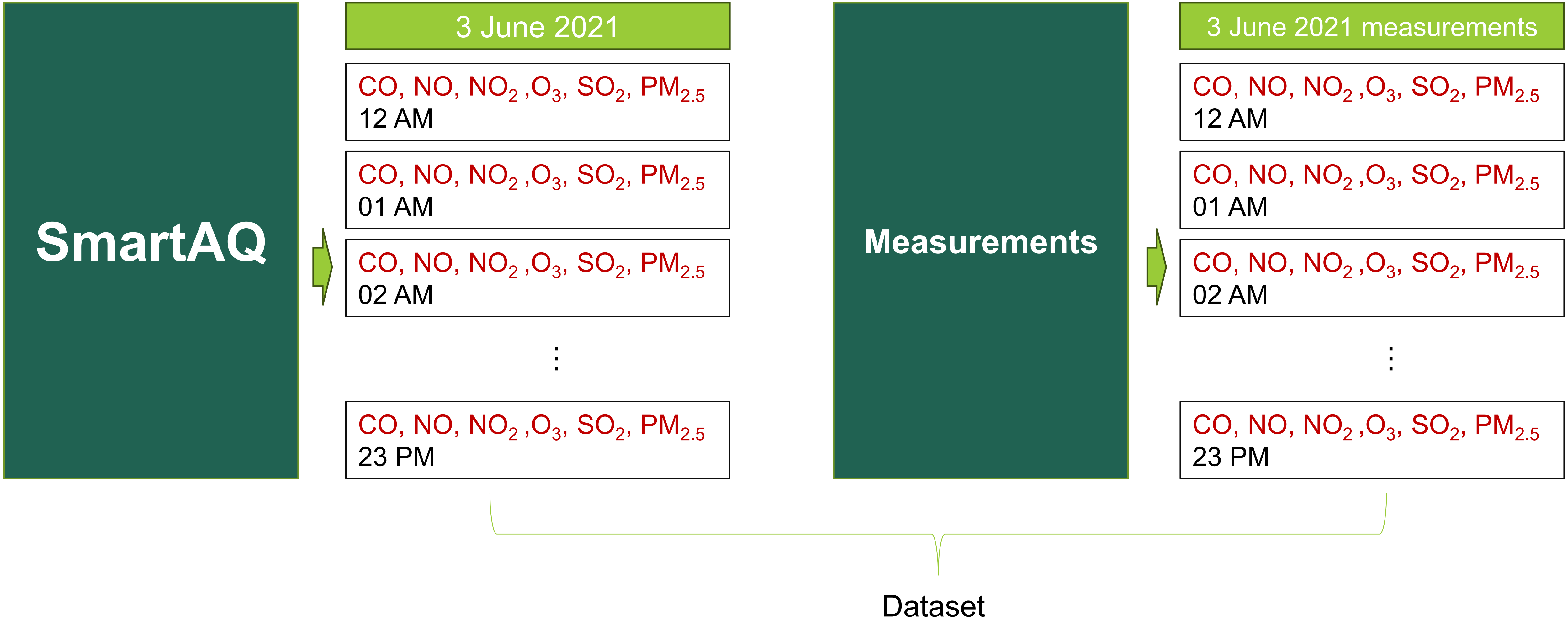
SmartAQ

Today' prediction	Tomorrow's prediction	The day after tomorrow
CO, NO, NO_2 ,O_3, SO_2, $PM_{2.5}$ 12 AM	CO, NO, NO_2 ,O_3, SO_2, $PM_{2.5}$ 12 AM	CO, NO, NO_2 ,O_3, SO_2, $PM_{2.5}$ 12 AM
CO, NO, NO_2 ,O_3, SO_2, $PM_{2.5}$ 01 AM	CO, NO, NO_2 ,O_3, SO_2, $PM_{2.5}$ 01 AM	CO, NO, NO_2 ,O_3, SO_2, $PM_{2.5}$ 01 AM
CO, NO, NO_2 ,O_3, SO_2, $PM_{2.5}$ 02 AM	CO, NO, NO_2 ,O_3, SO_2, $PM_{2.5}$ 02 AM	CO, NO, NO_2 ,O_3, SO_2, $PM_{2.5}$ 02 AM
⋮	⋮	⋮
CO, NO, NO_2 ,O_3, SO_2, $PM_{2.5}$ 23 PM	CO, NO, NO_2 ,O_3, SO_2, $PM_{2.5}$ 23 PM	CO, NO, NO_2 ,O_3, SO_2, $PM_{2.5}$ 23 PM
We keep this	We discard these	

Runs every day at
12AM – ends
around 14PM

SmartAQ operation – creating the training set for a point

Supposing that SmartAQ ran on 3 June 2021..
(started at 3 June 2021, 00:00 – ended 14:00)



SmartAQ

3 June 2021

CO, NO, NO_2 ,O_3, SO_2, $PM_{2.5}$
12 AM

CO, NO, NO_2 ,O_3, SO_2, $PM_{2.5}$
01 AM

CO, NO, NO_2 ,O_3, SO_2, $PM_{2.5}$
02 AM

⋮

CO, NO, NO_2 ,O_3, SO_2, $PM_{2.5}$
23 PM

Measurements

3 June 2021 measurements

CO, NO, NO_2 ,O_3, SO_2, $PM_{2.5}$
12 AM

CO, NO, NO_2 ,O_3, SO_2, $PM_{2.5}$
01 AM

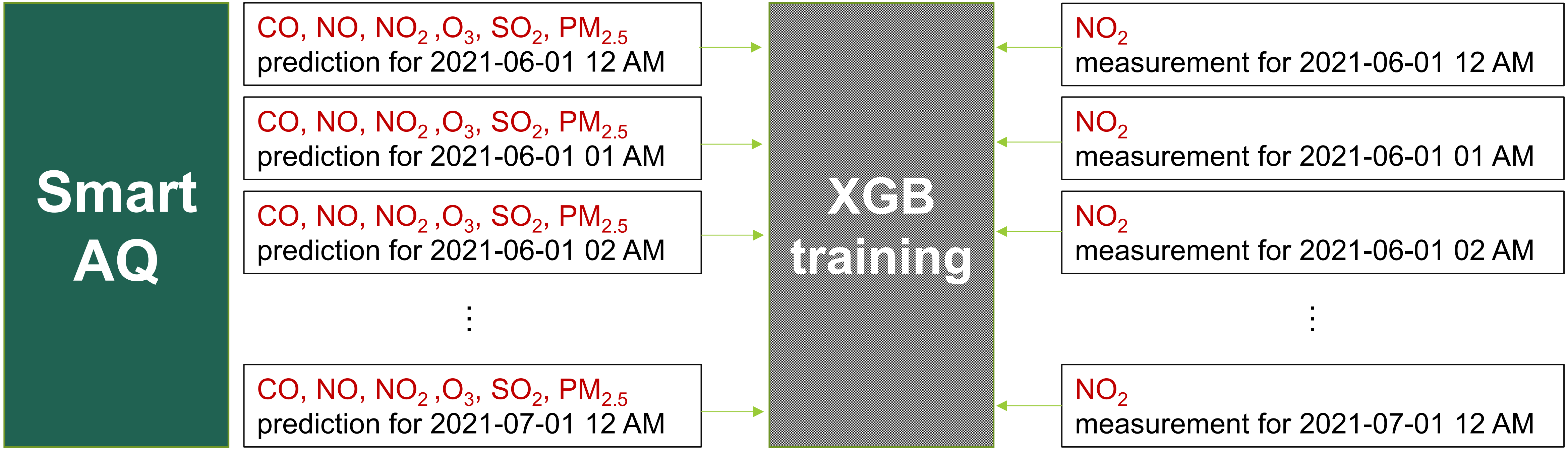
CO, NO, NO_2 ,O_3, SO_2, $PM_{2.5}$
02 AM

⋮

CO, NO, NO_2 ,O_3, SO_2, $PM_{2.5}$
23 PM

Dataset

XGBoost training



Smart AQ

$CO, NO, NO_2, O_3, SO_2, PM_{2.5}$ prediction for 2021-06-01 12 AM

$CO, NO, NO_2, O_3, SO_2, PM_{2.5}$ prediction for 2021-06-01 01 AM

$CO, NO, NO_2, O_3, SO_2, PM_{2.5}$ prediction for 2021-06-01 02 AM

⋮

$CO, NO, NO_2, O_3, SO_2, PM_{2.5}$ prediction for 2021-07-01 12 AM

XGB training

NO_2 measurement for 2021-06-01 12 AM

NO_2 measurement for 2021-06-01 01 AM

NO_2 measurement for 2021-06-01 02 AM

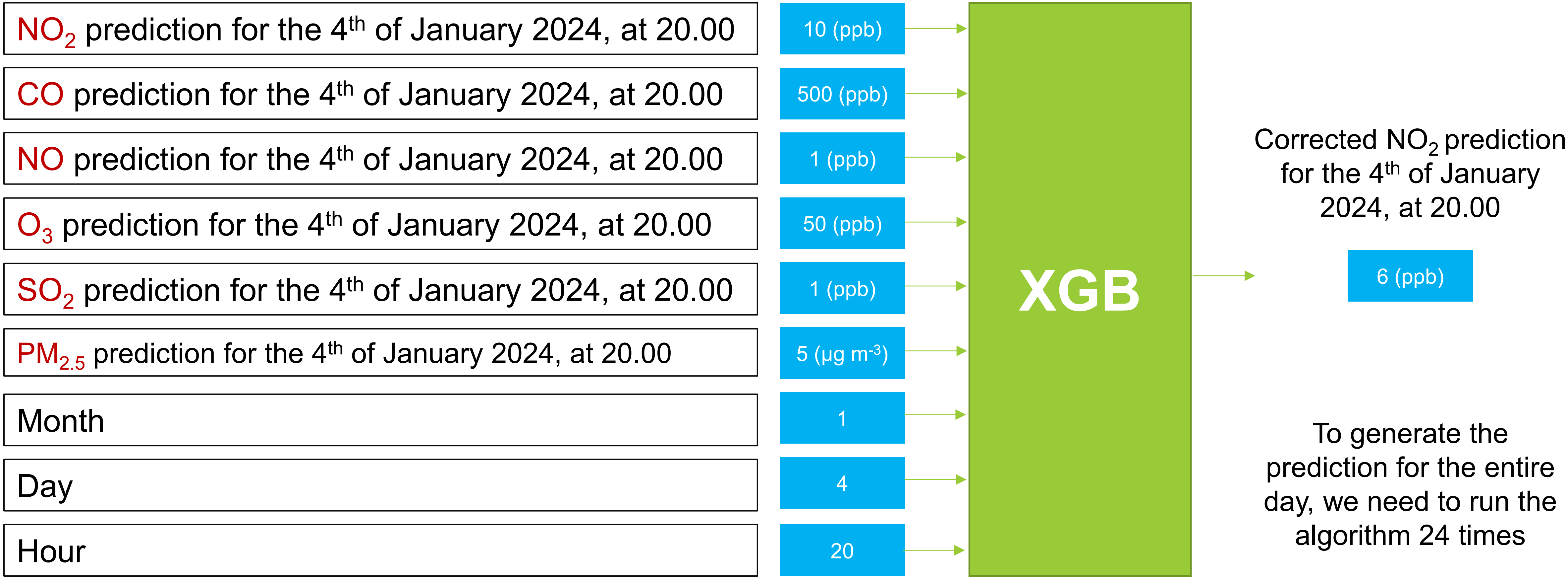
⋮

NO_2 measurement for 2021-07-01 12 AM

- Hourly predictions and measurements for **the entire training period**
- Learns to change the prediction for each hour so that it minimizes the MSE, computed according to the measurements of each hour

Inputs and Outputs of a trained XGBoost model

Supposing that we want to have the corrected NO_2 for the 4th of January 2024, at 20.00



NO_2 prediction for the 4th of January 2024, at 20.00	10 (ppb)
CO prediction for the 4th of January 2024, at 20.00	500 (ppb)
NO prediction for the 4th of January 2024, at 20.00	1 (ppb)
O_3 prediction for the 4th of January 2024, at 20.00	50 (ppb)
SO_2 prediction for the 4th of January 2024, at 20.00	1 (ppb)
$PM_{2.5}$ prediction for the 4th of January 2024, at 20.00	5 (μg m^{-3})
Month	1
Day	4
Hour	20

Corrected NO_2 prediction for the 4th of January 2024, at 20.00

6 (ppb)

To generate the prediction for the entire day, we need to run the algorithm 24 times

XGBoost training

Let’s build a Decision Tree first

CO, NO, NO_2, O_3
CO, NO, NO_2, O_3
CO, NO, NO_2, O_3
CO, NO, NO_2, O_3

Raw data

NO_2 = 10 ppb
NO_2 = 15 ppb
NO_2 = 12 ppb
NO_2 = 11 ppb

Measurement

Mean Measured NO_2 = 12

Split the data into two groups based on a feature

CO, NO, NO_2, O_3
CO, NO, NO_2, O_3

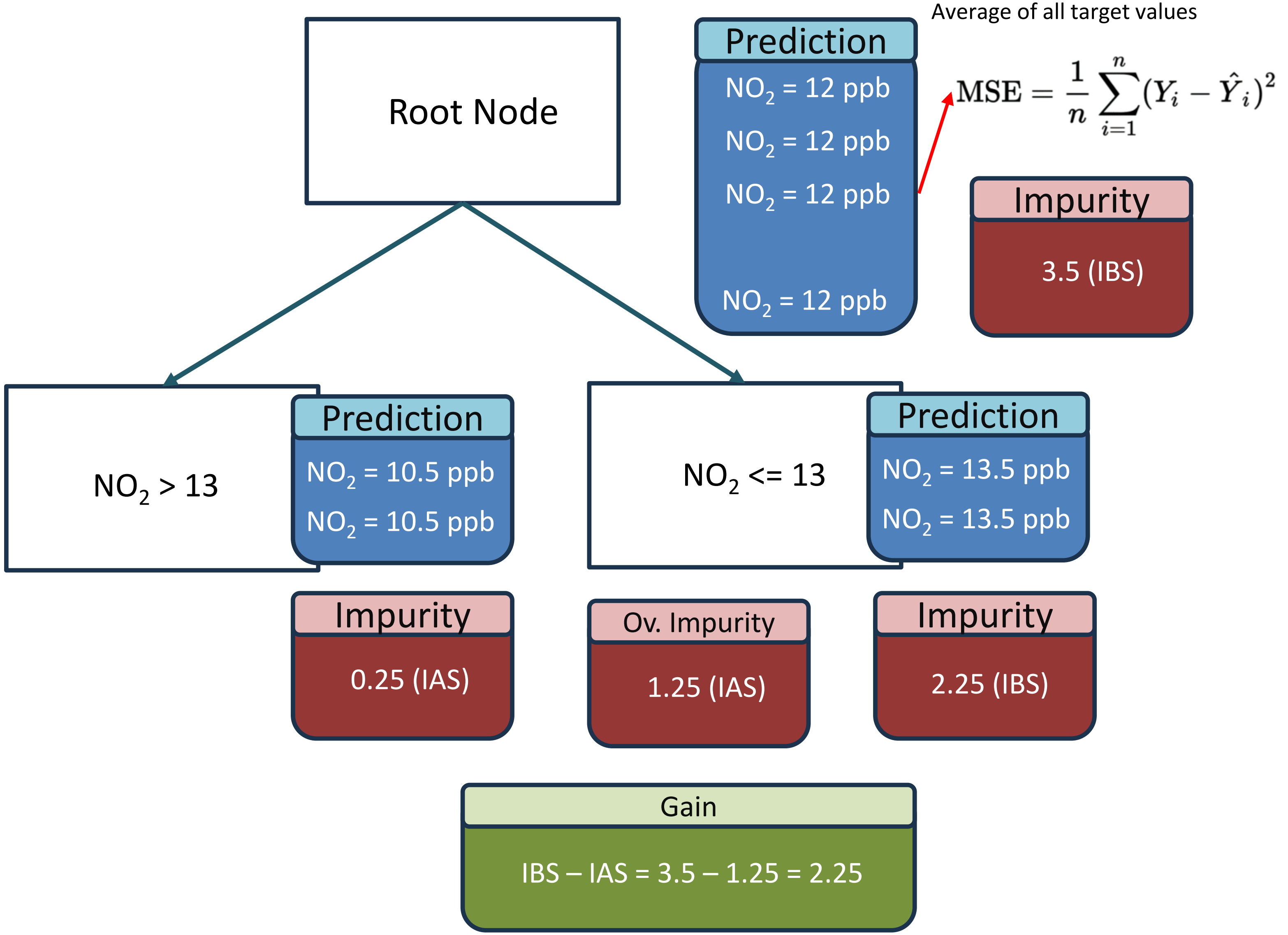
NO_2 = 10 ppb
NO_2 = 11 ppb

NO_2 > 13

CO, NO, NO_2, O_3
CO, NO, NO_2, O_3

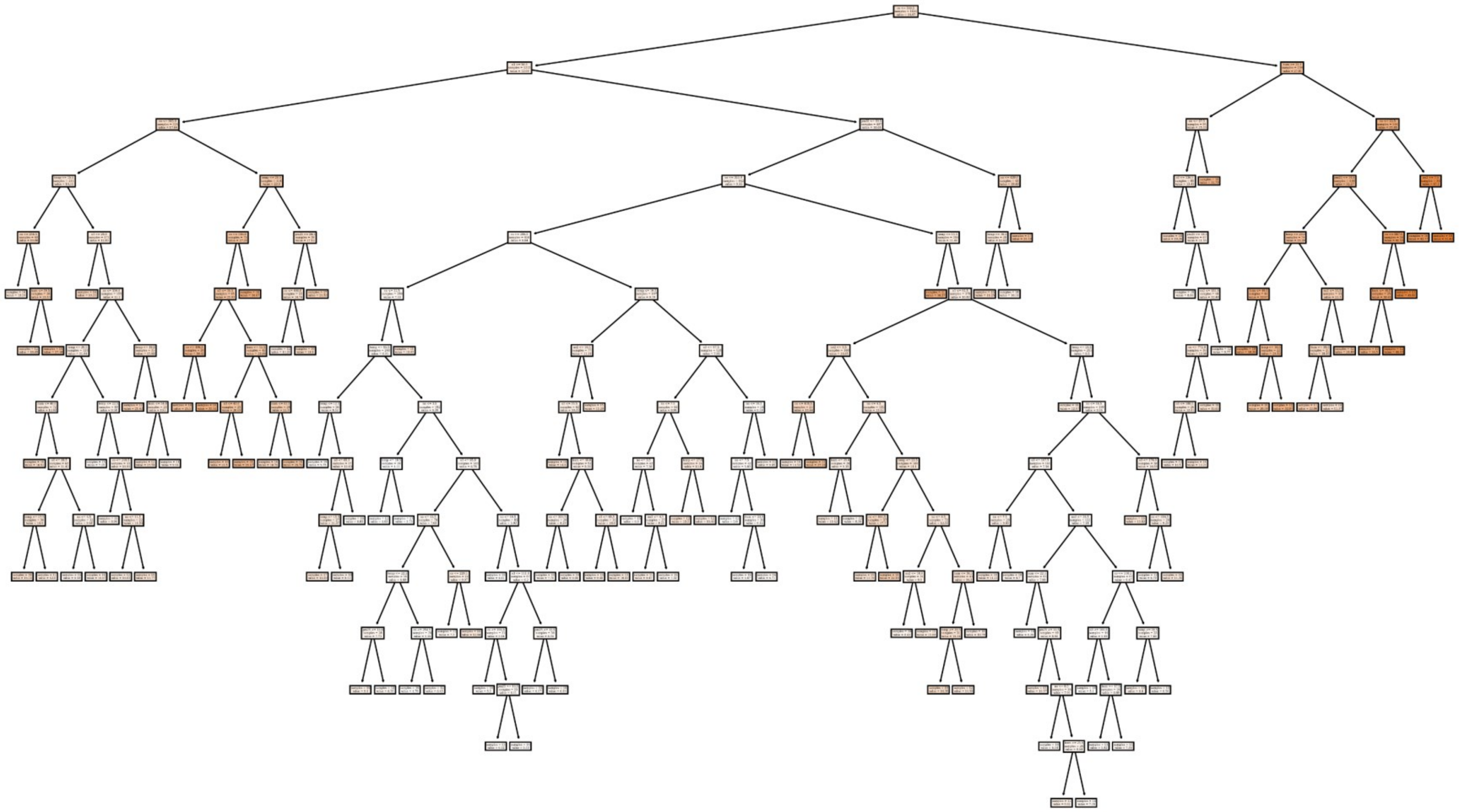
NO_2 = 15 ppb
NO_2 = 12 ppb

NO_2 <= 13



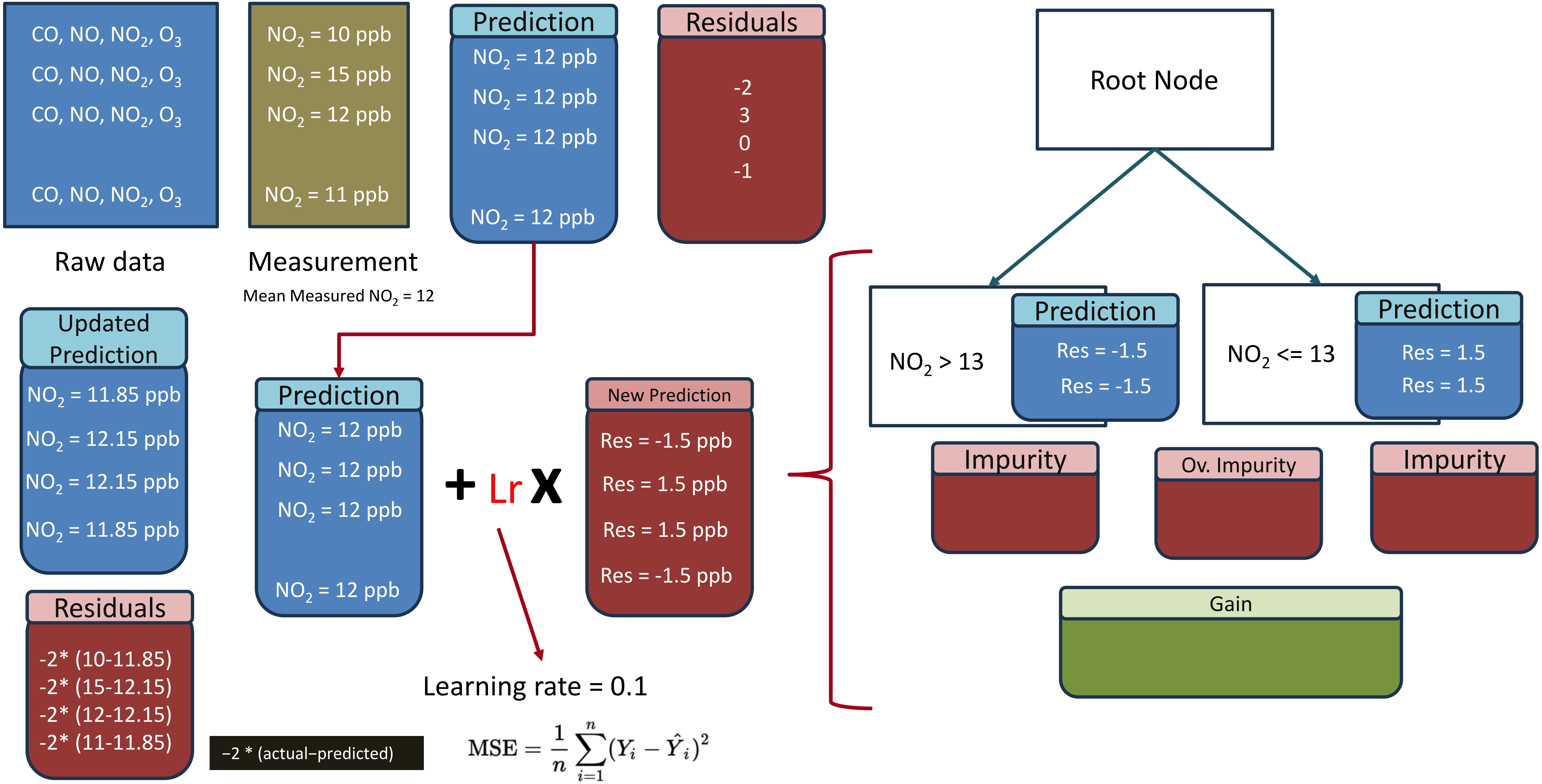
Average of all target values

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$



Typical 13-level tree

The first tree of the XGBoost model



CO, NO, NO_2, O_3
CO, NO, NO_2, O_3
CO, NO, NO_2, O_3
CO, NO, NO_2, O_3

Raw data

NO_2 = 10 ppb
NO_2 = 15 ppb
NO_2 = 12 ppb
NO_2 = 11 ppb

Measurement

Mean Measured NO_2 = 12

Prediction
NO_2 = 12 ppb
NO_2 = 12 ppb
NO_2 = 12 ppb
NO_2 = 12 ppb

Residuals
-2
3
0
-1

Root Node

NO_2 > 13

Prediction
Res = -1.5
Res = -1.5

NO_2 <= 13

Prediction
Res = 1.5
Res = 1.5

Impurity

Ov. Impurity

Impurity

Gain

Updated Prediction
NO_2 = 11.85 ppb
NO_2 = 12.15 ppb
NO_2 = 12.15 ppb
NO_2 = 11.85 ppb

Residuals
-2* (10-11.85)
-2* (15-12.15)
-2* (12-12.15)
-2* (11-11.85)

−2 * (actual−predicted)

Prediction
NO_2 = 12 ppb
NO_2 = 12 ppb
NO_2 = 12 ppb
NO_2 = 12 ppb

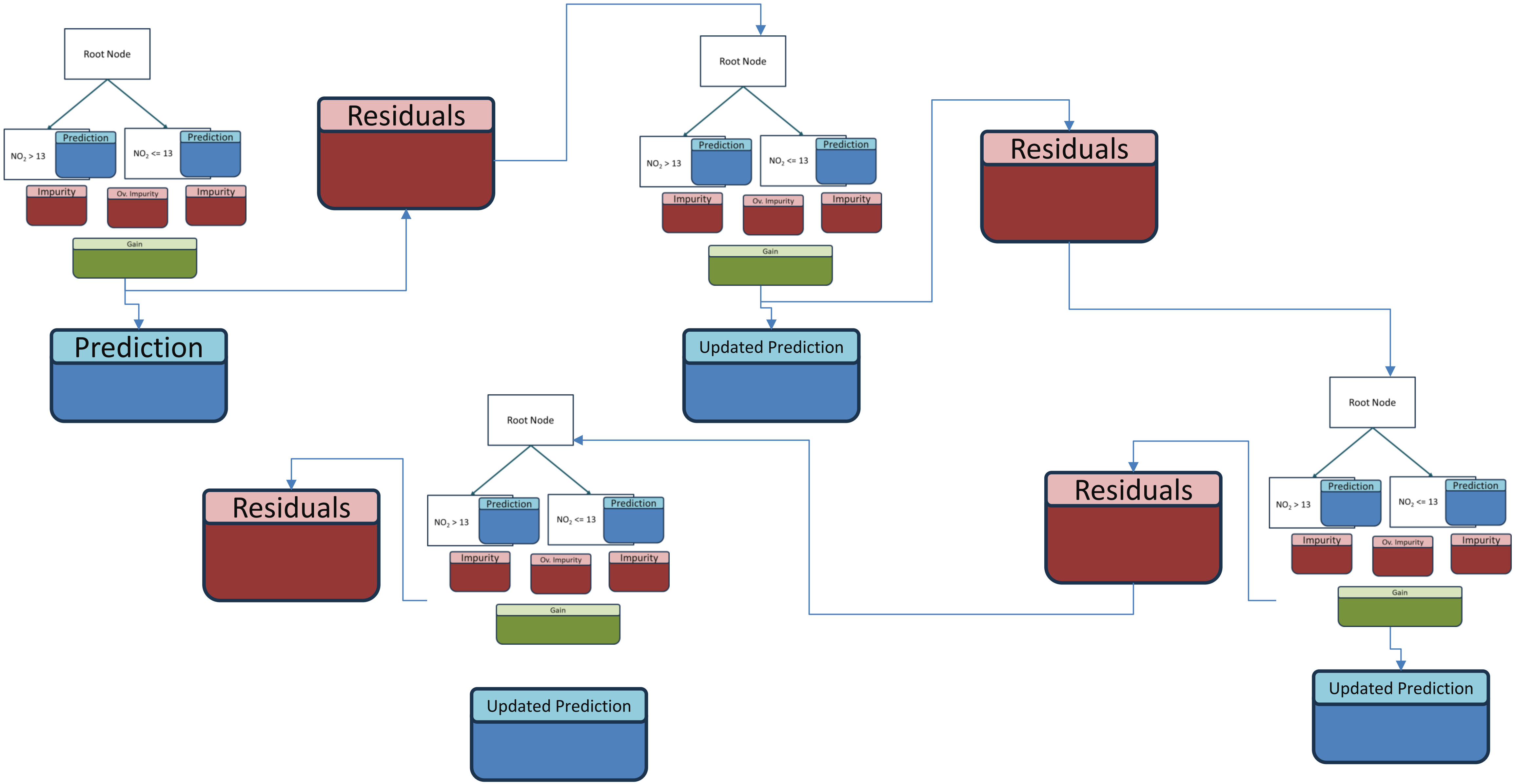
+ Lr X

New Prediction
Res = -1.5 ppb
Res = 1.5 ppb
Res = 1.5 ppb
Res = -1.5 ppb

Learning rate = 0.1

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

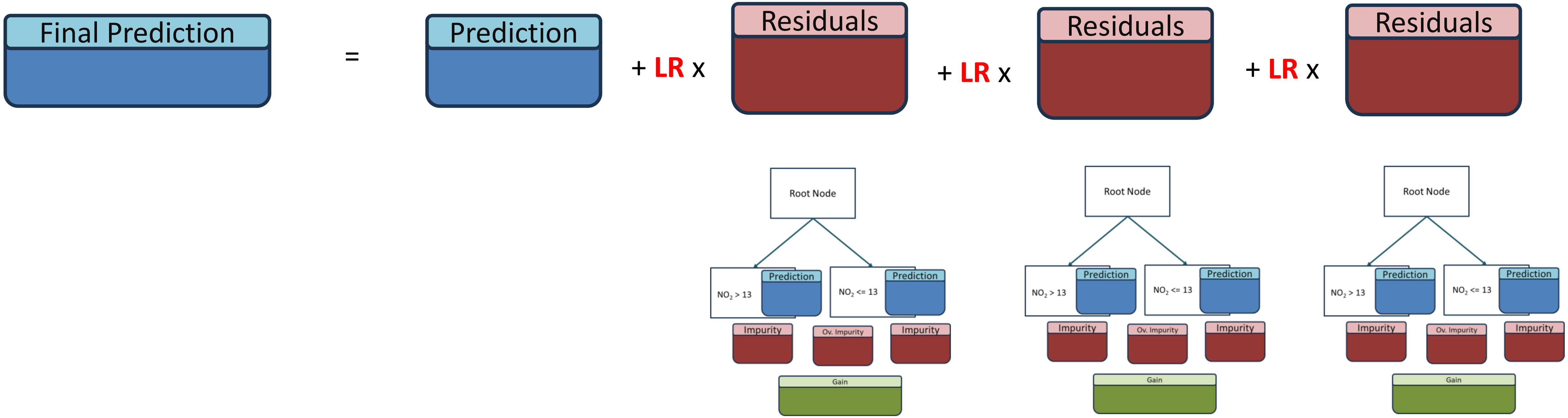
The "forest" of the XGBoost model



The ultimate prediction function

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

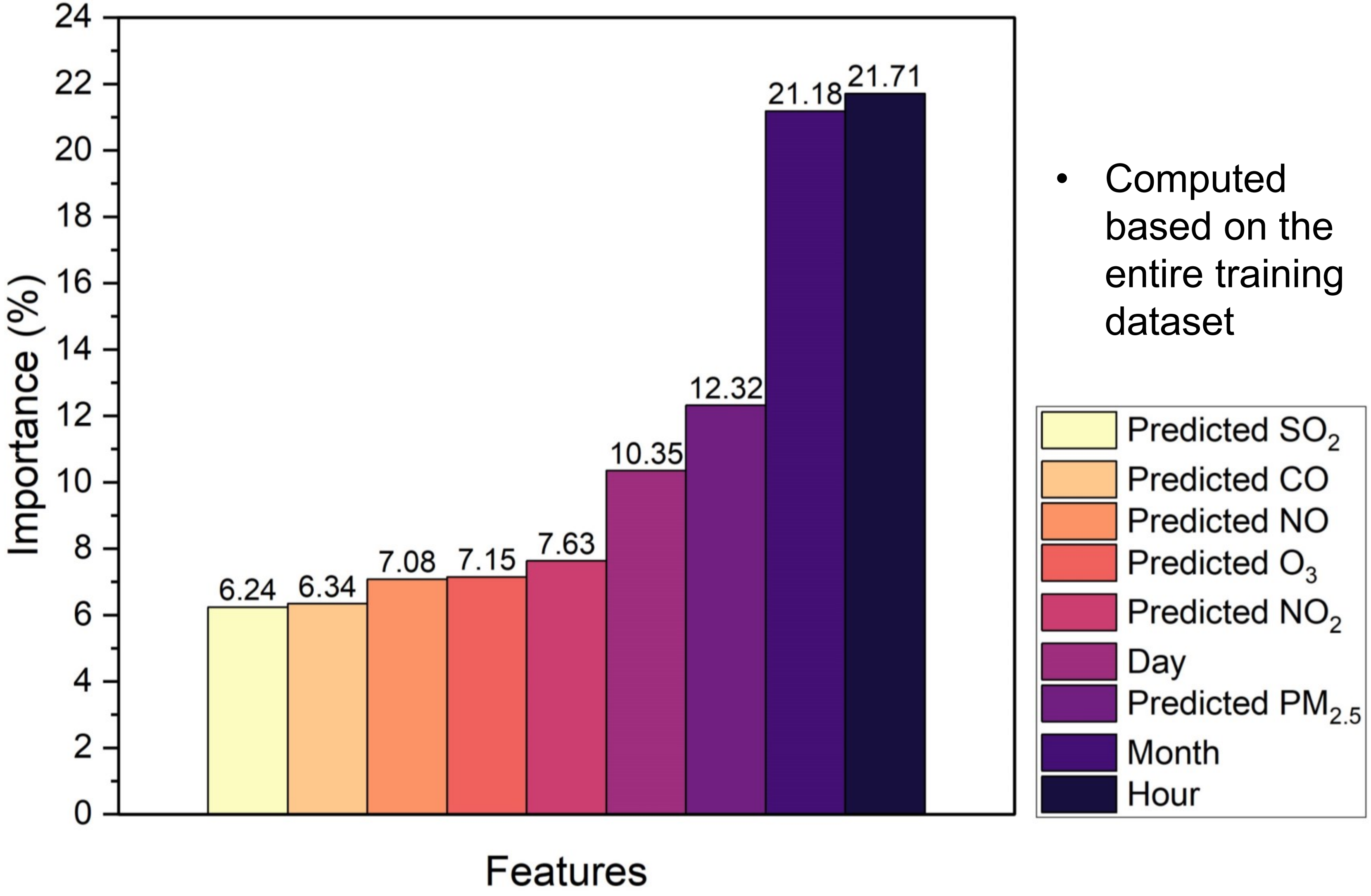
$$-2 \times (Y_{true} - Y_{pred})$$



Final Prediction = Prediction + **LR** x Residuals + **LR** x Residuals + **LR** x Residuals

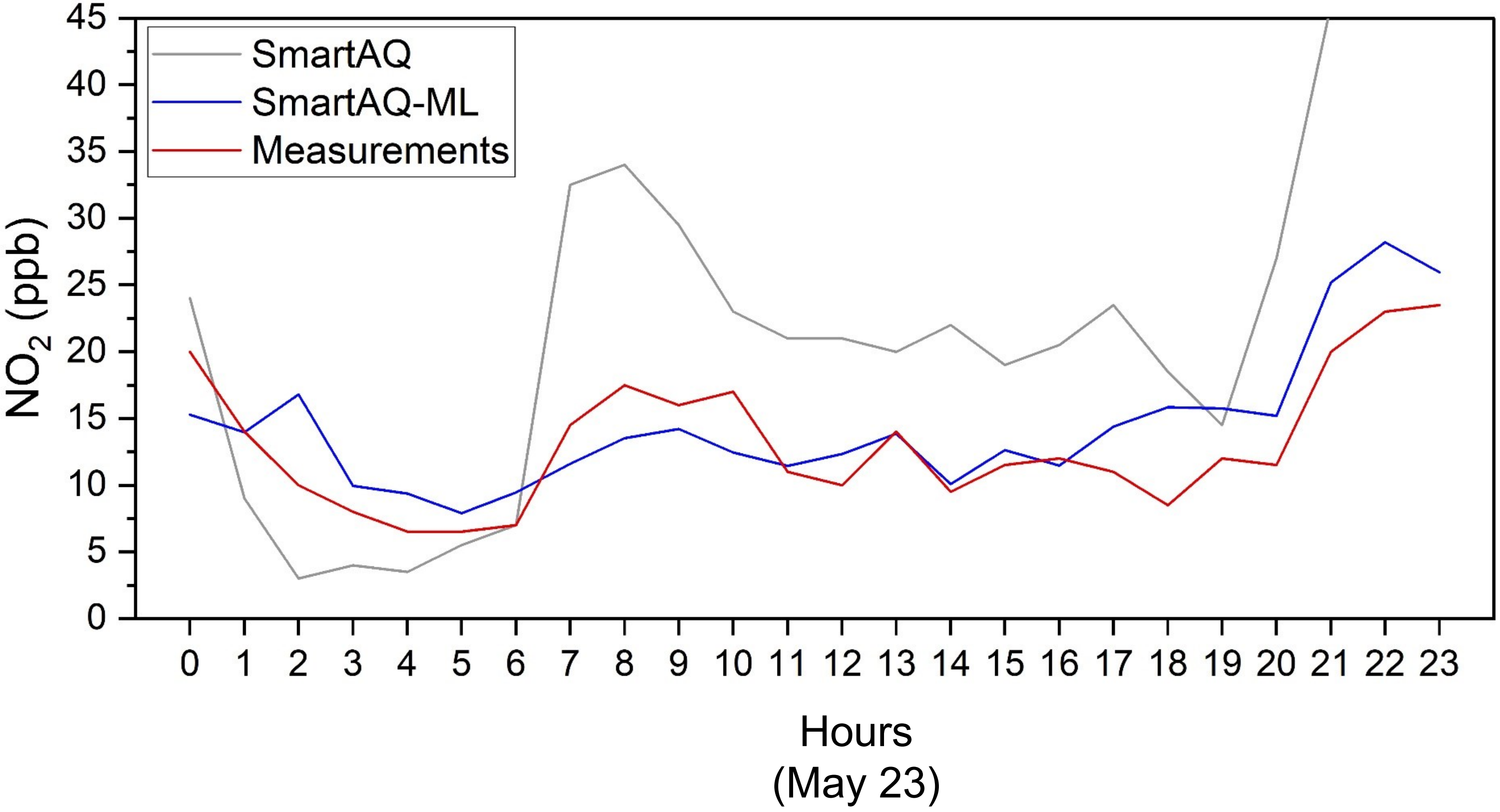
Results
(test period)

Feature Importance based on XGBoost



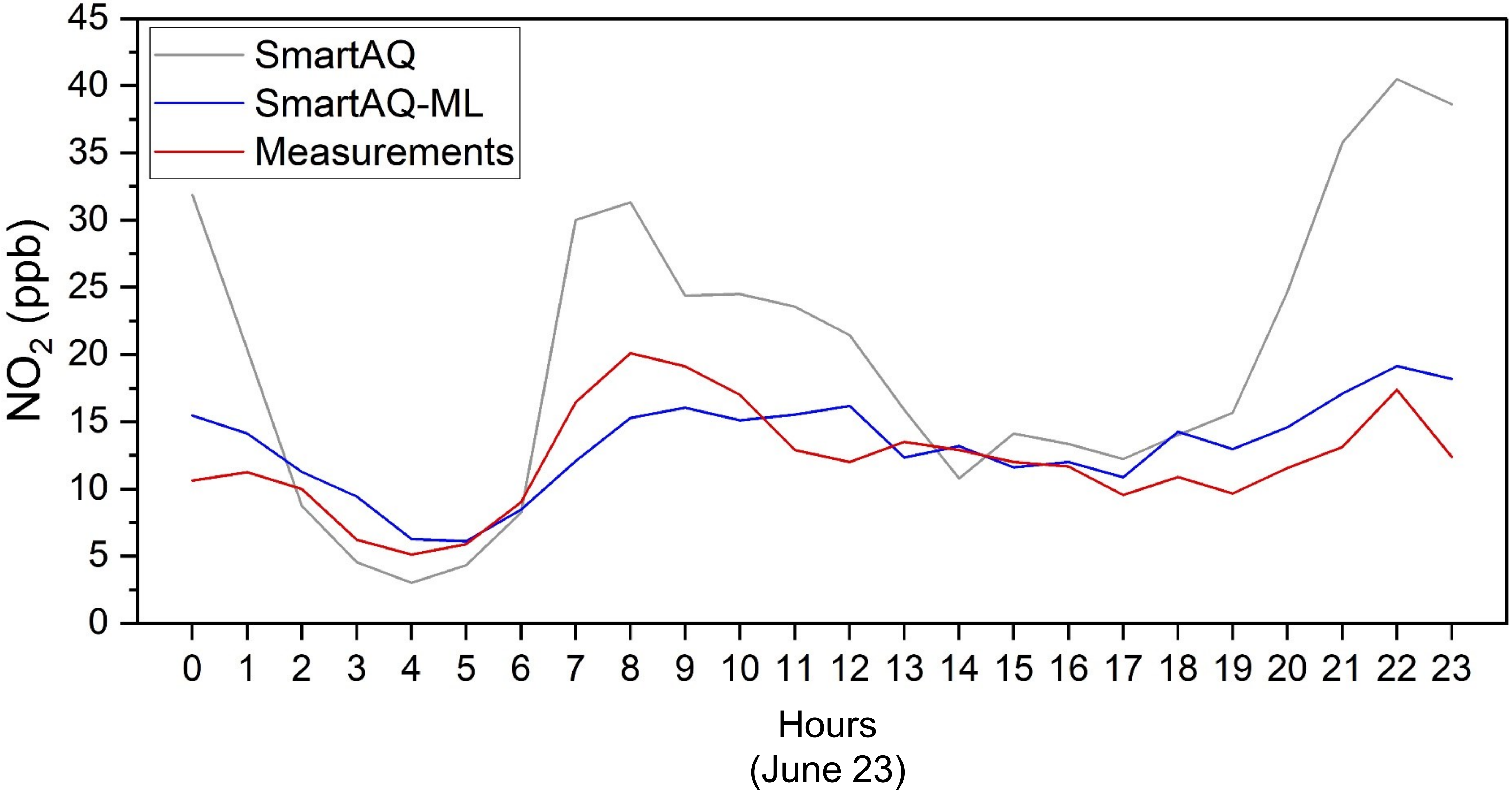
- Computed based on the entire training dataset

NO_2 average diurnal variation May 2023

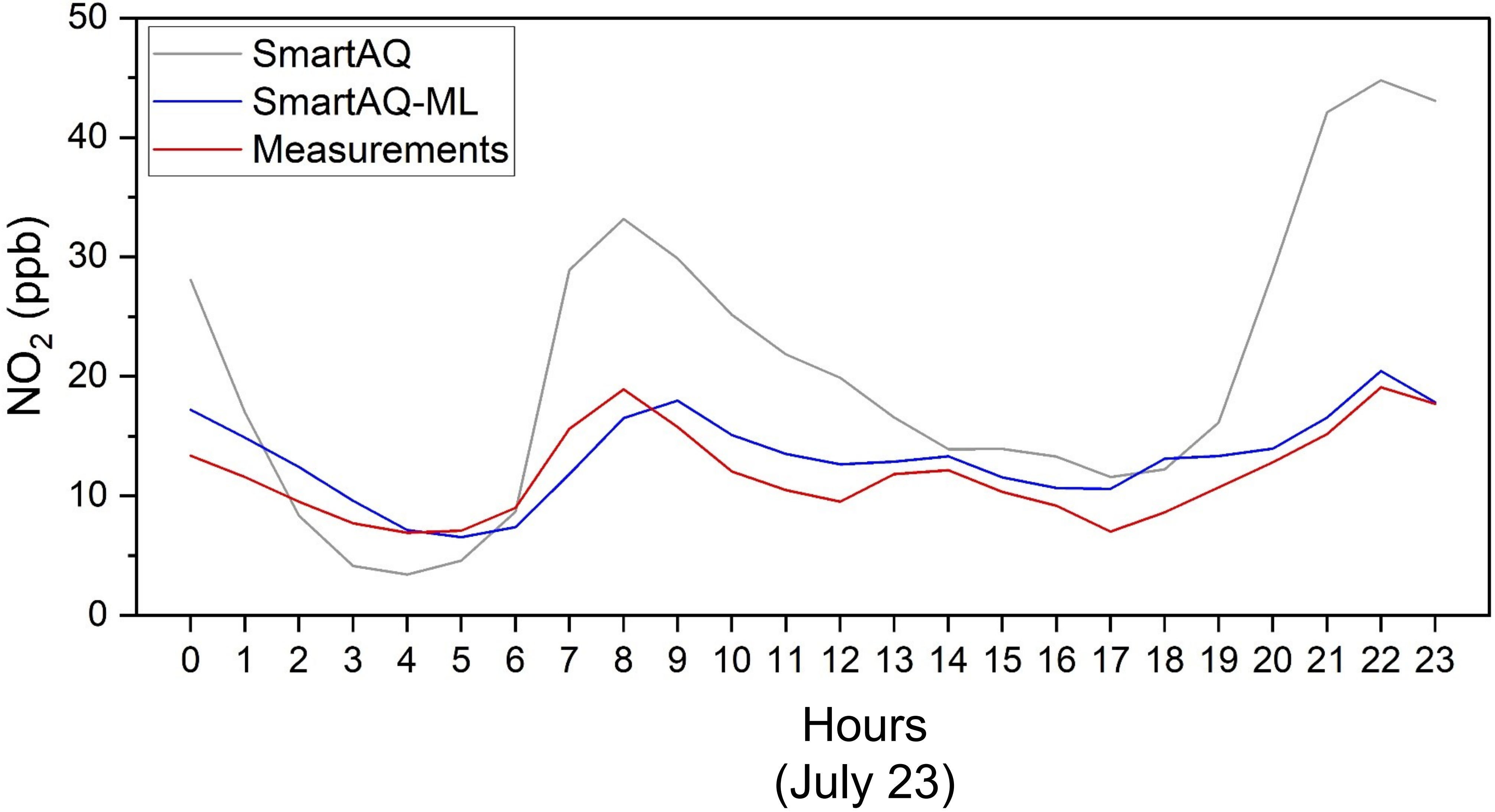


- Under-estimation during the night
- Over-estimation during the day
- Effective correction of ML

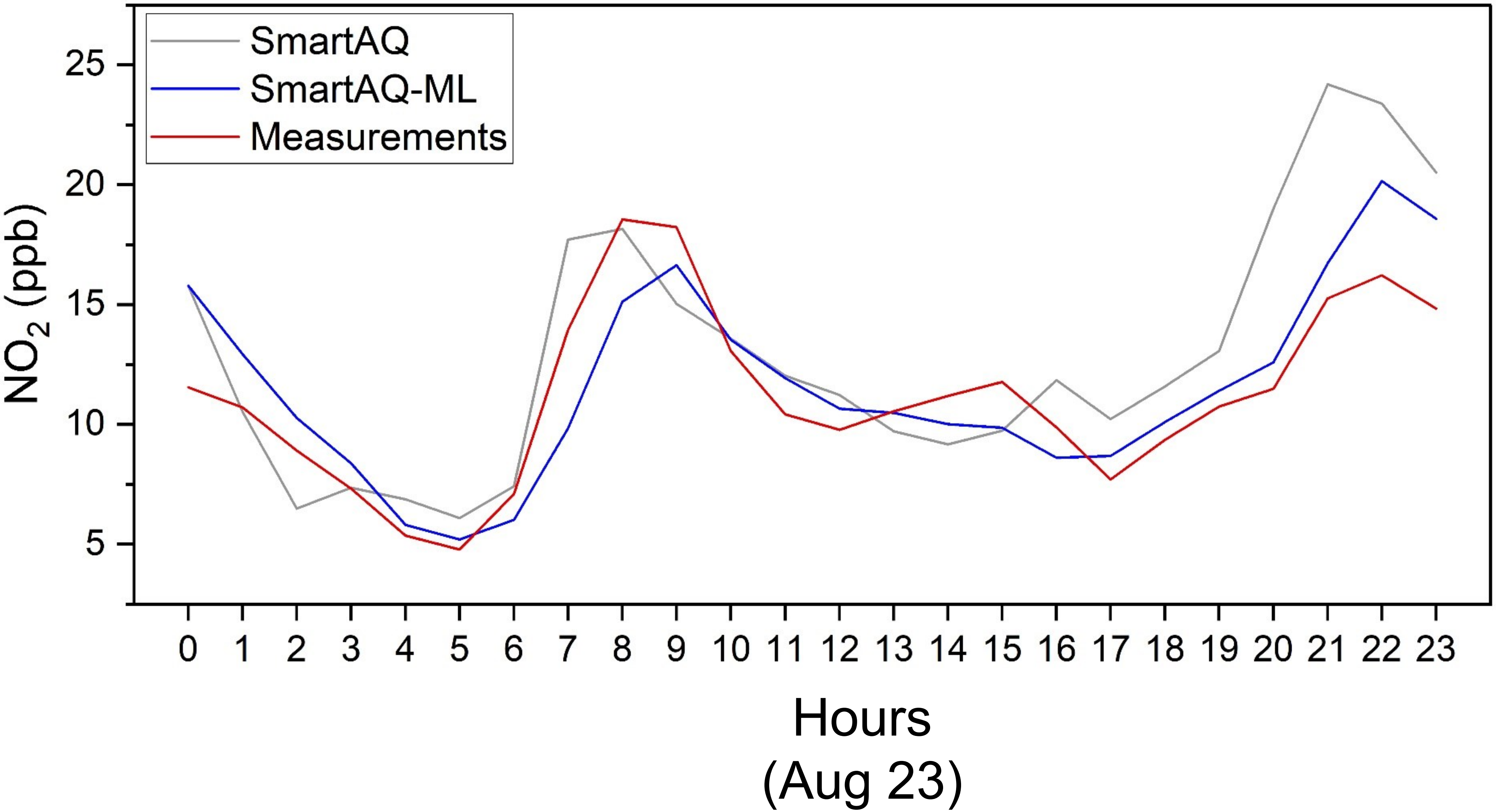
NO_2 average diurnal variation June 2023



NO_2 average diurnal variation July 2023

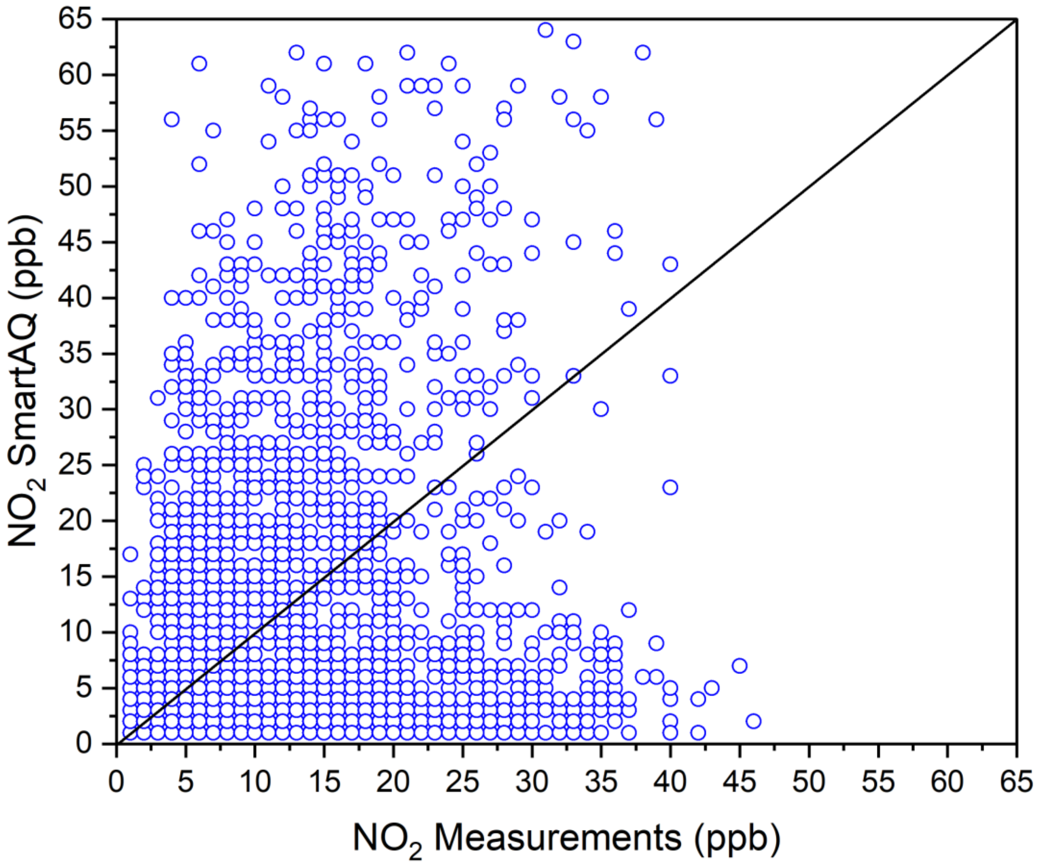


NO_2 average diurnal variation August 2023

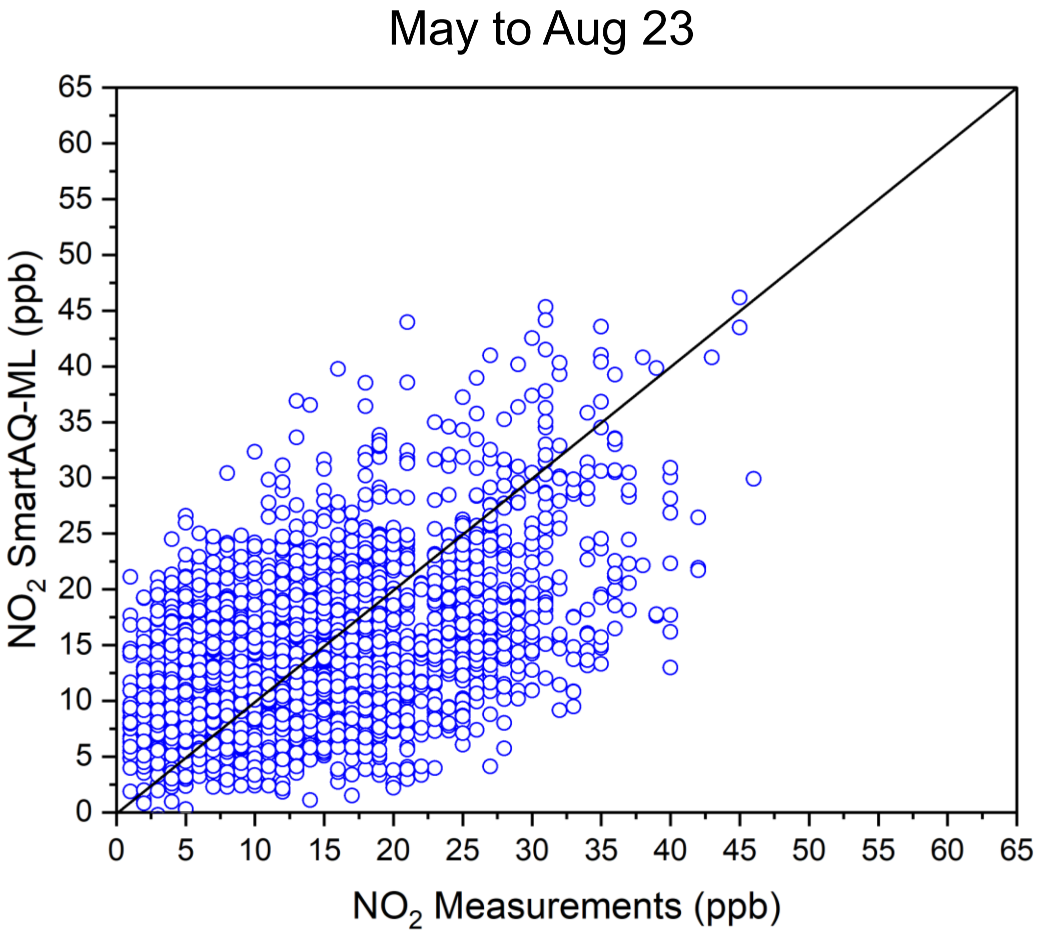


Overall scatter plot - SmartAQ

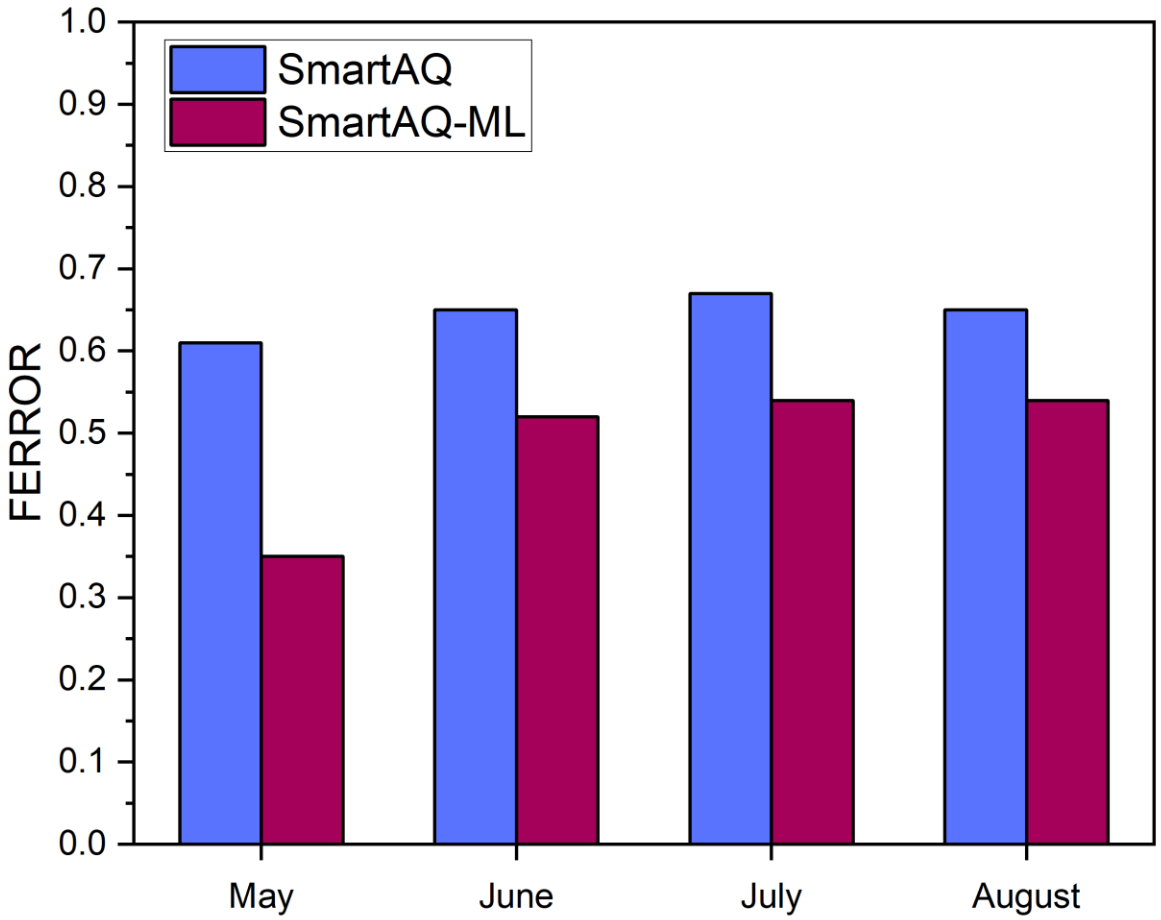
May to Aug 23



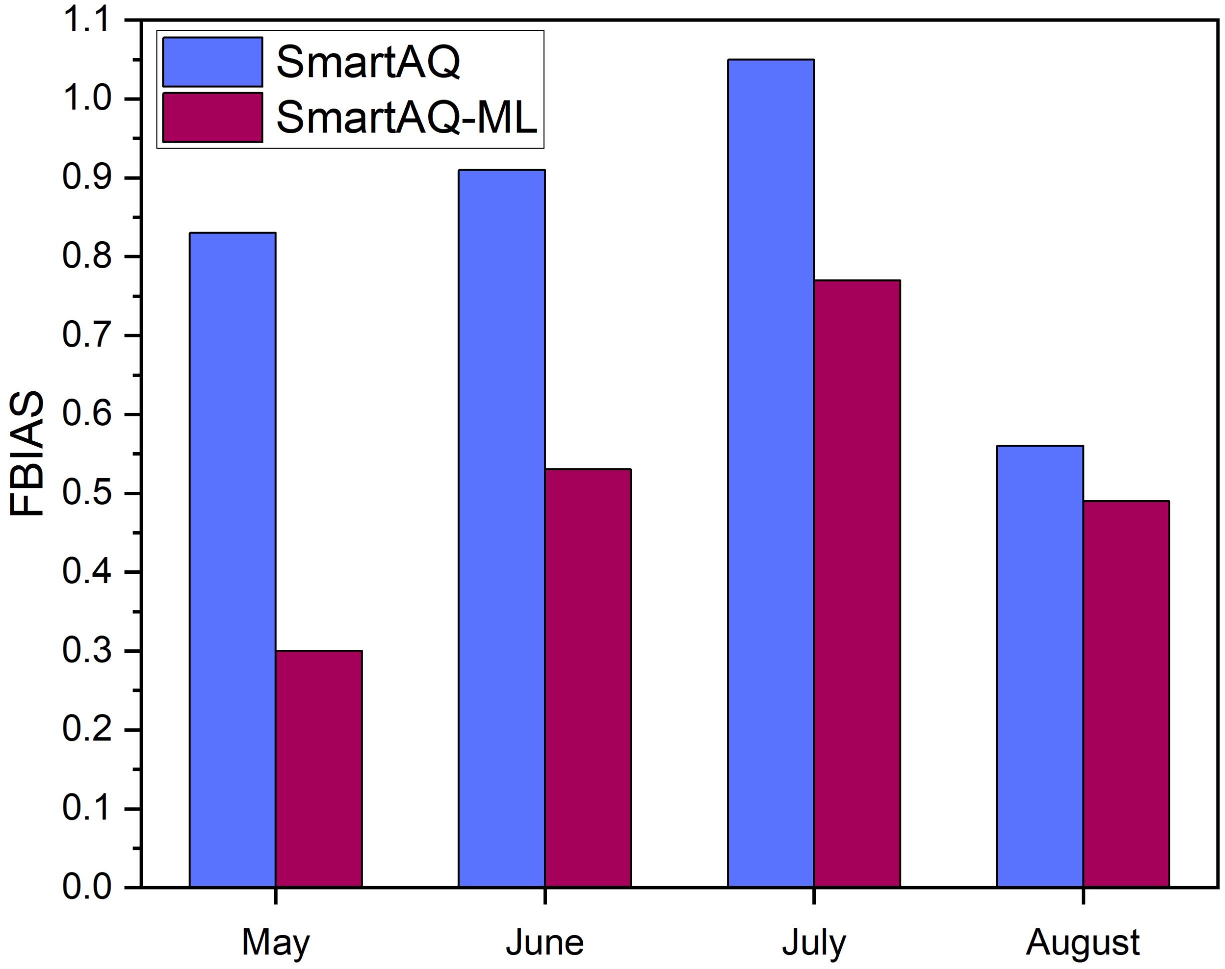
Overall scatter plot – SmartAQ-ML



FERROR comparison by month



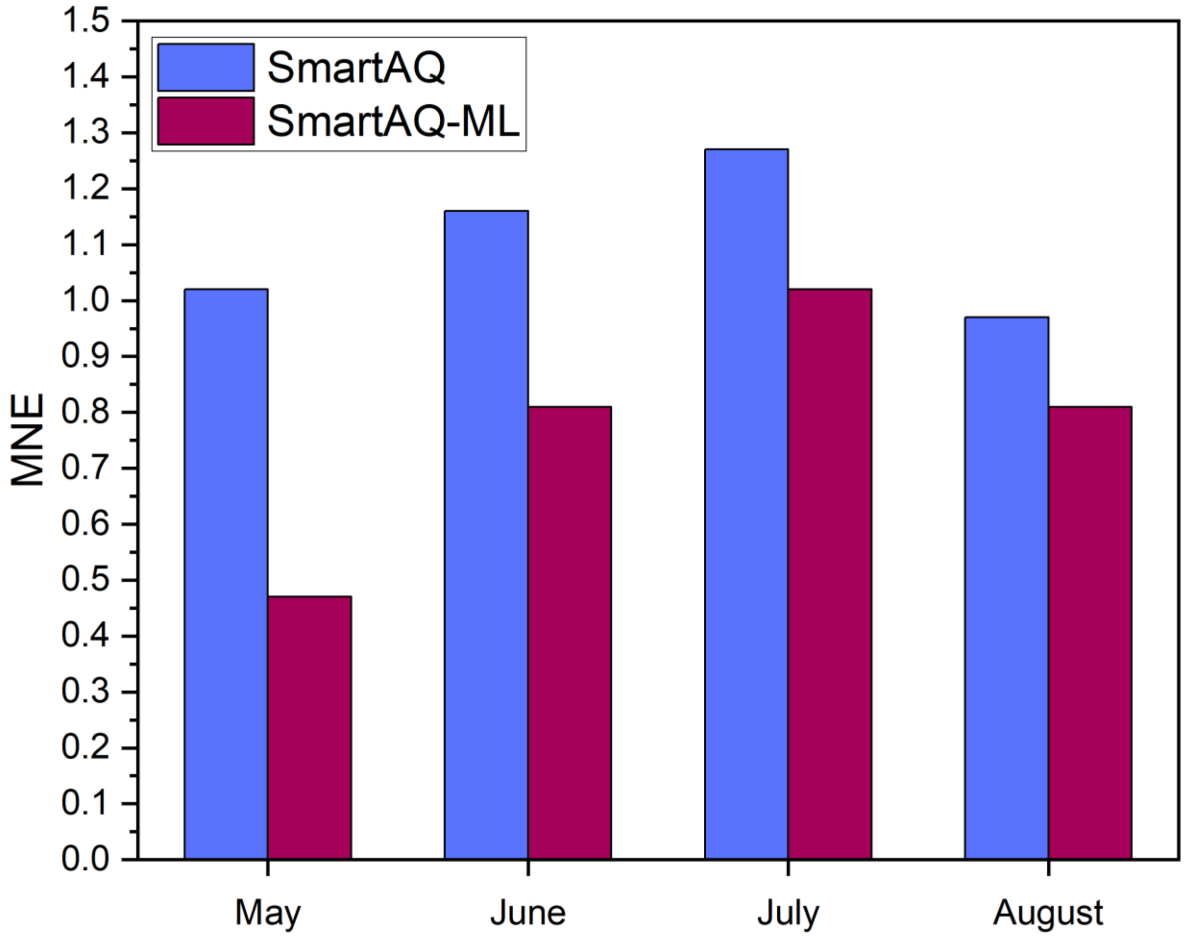
FBIAS comparison by month



MNE comparison by month

Metrics

- Significant reduction of FBIAS
- ~23% reduction of FERROR
- ~35% reduction of MNE



Next steps

- Inspect how the algorithm works in the entire grid of the greater Patras area
- Attempt to explain potential mistakes
- Attempt to improve the correction further using more input features (sensors, meteorology)
- Test the methodology at the National Observatory of Athens (when the SmartAQ is ready)
- Benchmark more ML algorithms (e.g. LSTM, Transformers)
- Use the methodology to improve the forecast of other pollutants

Contributors

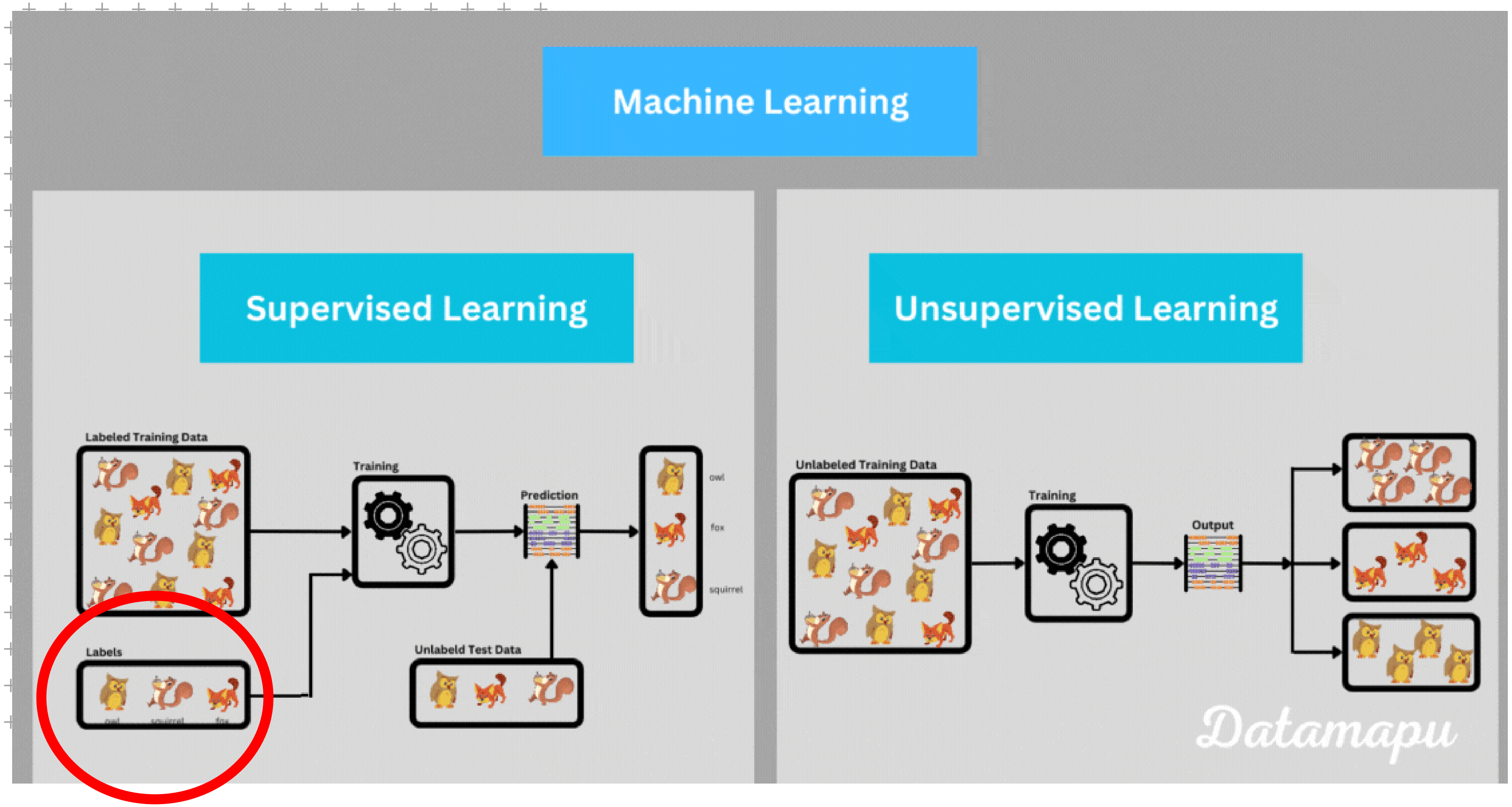
FORTH
Institute of Chemical Engineering Sciences

C·STACC
Center for the Study of Air Quality & Climate Change

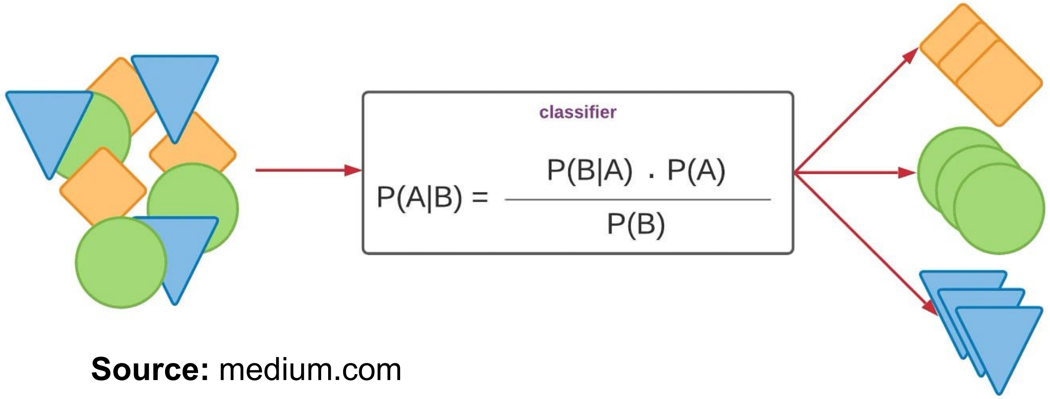
- Prof. Spyros Pandis
- Dr. George Fouskas
- Dr. Valia Siouti

Thank You!

Supervised and Unsupervised Machine Learning



Naive Bayes Classifier



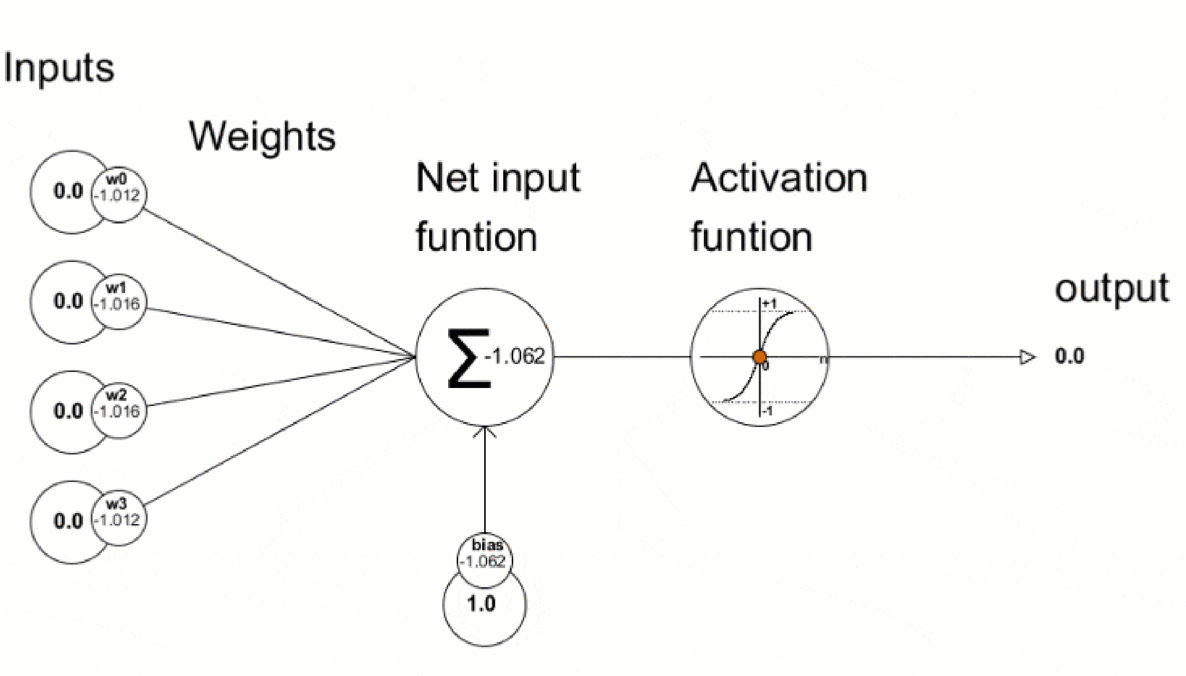
Source: medium.com

A is the class label
B is the set of features (a_1, a_2, a_3,…,a_n)

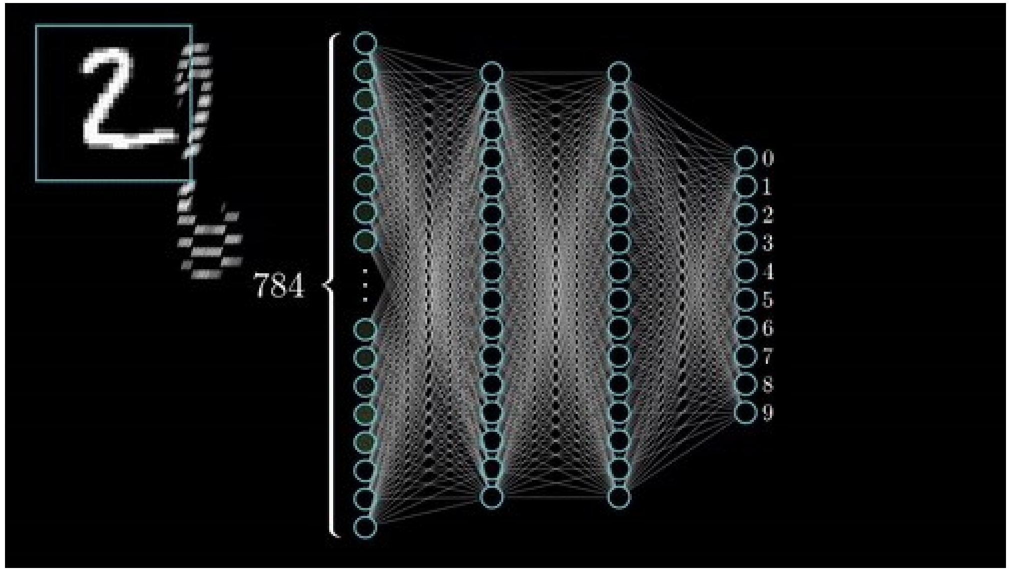
Everyone is aware that this algorithm is naive because it assumes that measurement features are independent of one another and contribute equally to the outcome.

The Neural Network

Introduction to the Neural Network



Source: www.medium.com



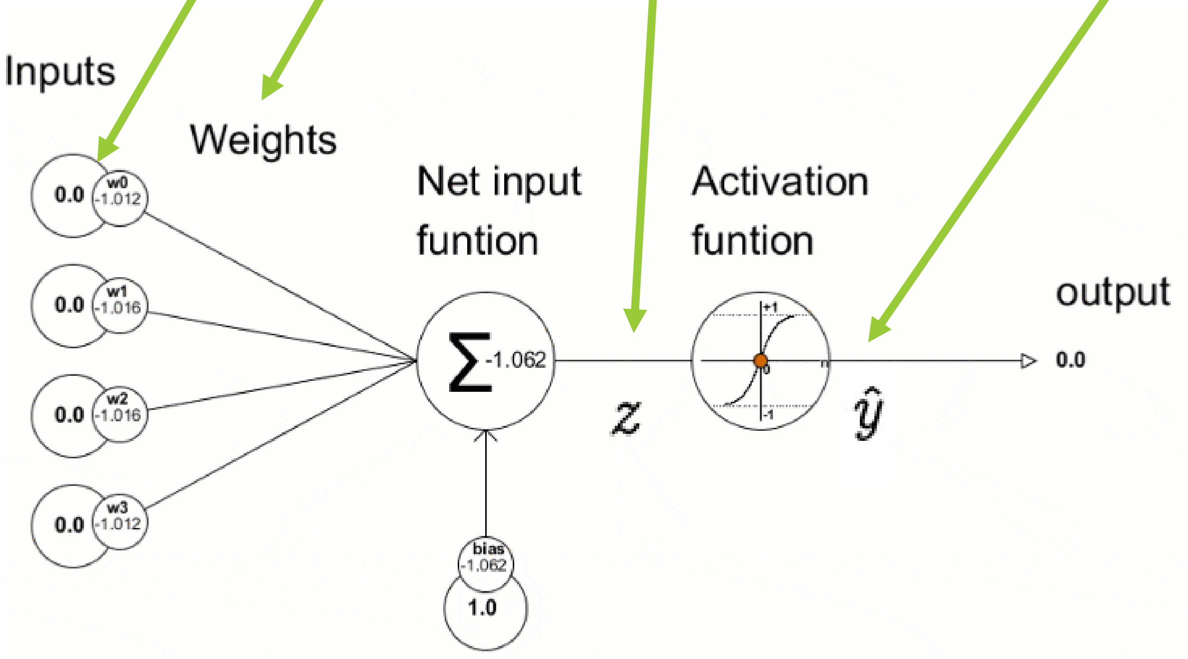
Source: www.medium.com

The forward pass

$$\sum = (x_1 \times w_1) + (x_2 \times w_2) + \cdots + (x_n \times w_n)$$

$$z = x.w + b$$

$$\hat{y} = \sigma(z) = \frac{1}{1 + e^{-z}}$$



Source: www.medium.com

The back-propagation

$$C = MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

→

$$\frac{\partial C}{\partial w_i} = \frac{\partial C}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial z} \times \frac{\partial z}{\partial w_i}$$

→

$$\frac{\partial C}{\partial \hat{y}} = ?$$

🟦

$$\frac{\partial \hat{y}}{\partial z} = ?$$

🟩

$$\frac{\partial z}{\partial w_1} = ?$$

🟨

🟩

$$\begin{aligned}
\frac{\partial \hat{y}}{\partial z} &= \frac{\partial}{\partial z}\sigma(z) \\
&= \frac{\partial}{\partial z}\left(\frac{1}{1+e^{-z}}\right) \\
&= \frac{e^{-z}}{(1+e^{-z})^2} \\
&= \frac{1}{(1+e^{-z})} \times \frac{e^{-z}}{(1+e^{-z})} \\
&= \frac{1}{(1+e^{-z})} \times \left(1 - \frac{1}{(1+e^{-z})}\right) \\
&= \sigma(z) \times (1 - \sigma(z))
\end{aligned}$$

🟦

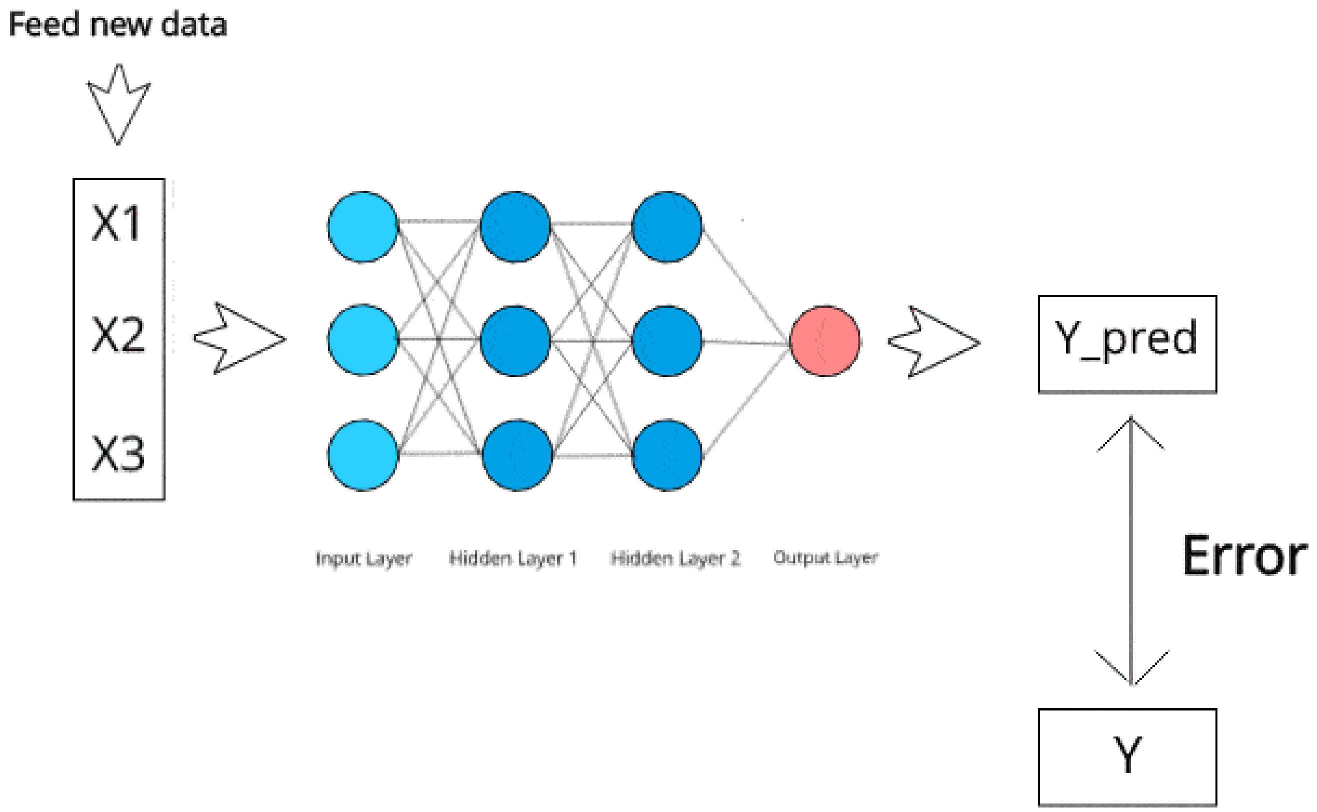
$$\frac{\partial C}{\partial \hat{y}} = \frac{2}{n} \times sum(y - \hat{y})$$

🟨

$$\frac{\partial C}{\partial w_i} = \frac{2}{n} \times sum(y - \hat{y}) \times \sigma(z) \times (1 - \sigma(z)) \times x_i$$

$$\frac{\partial C}{\partial b} = \frac{2}{n} \times sum(y - \hat{y}) \times \sigma(z) \times (1 - \sigma(z))$$

Summary and Optimization



Feed new data

X1

X2

X3

Input Layer

Hidden Layer 1

Hidden Layer 2

Output Layer

Y_pred

Error

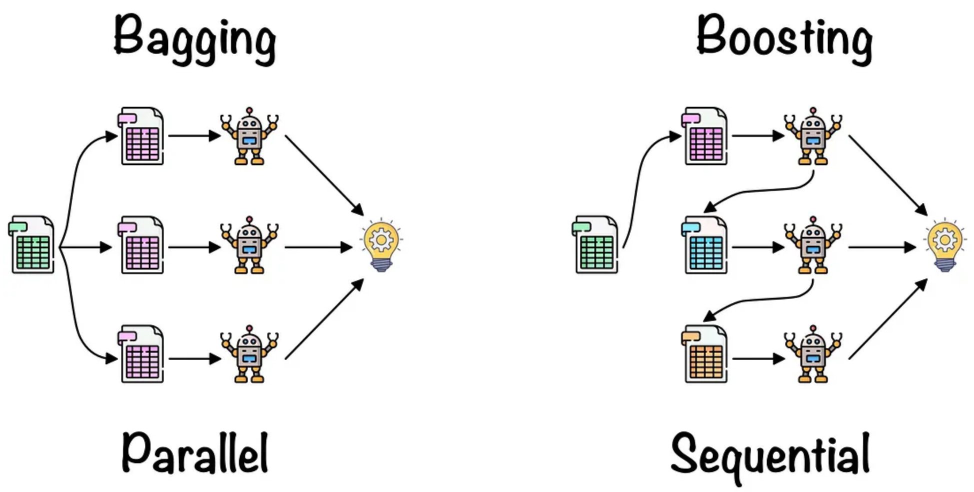
Y

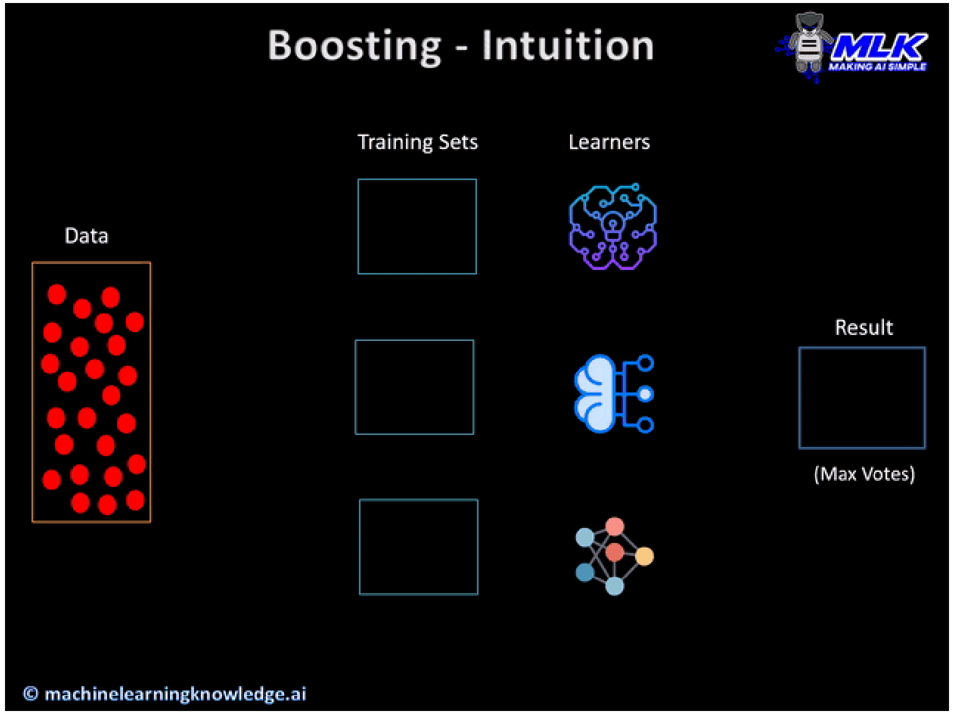
Gradient descent optimization

$$w_i = w_i - \left(\alpha \times \frac{\partial C}{\partial w_i}\right)$$

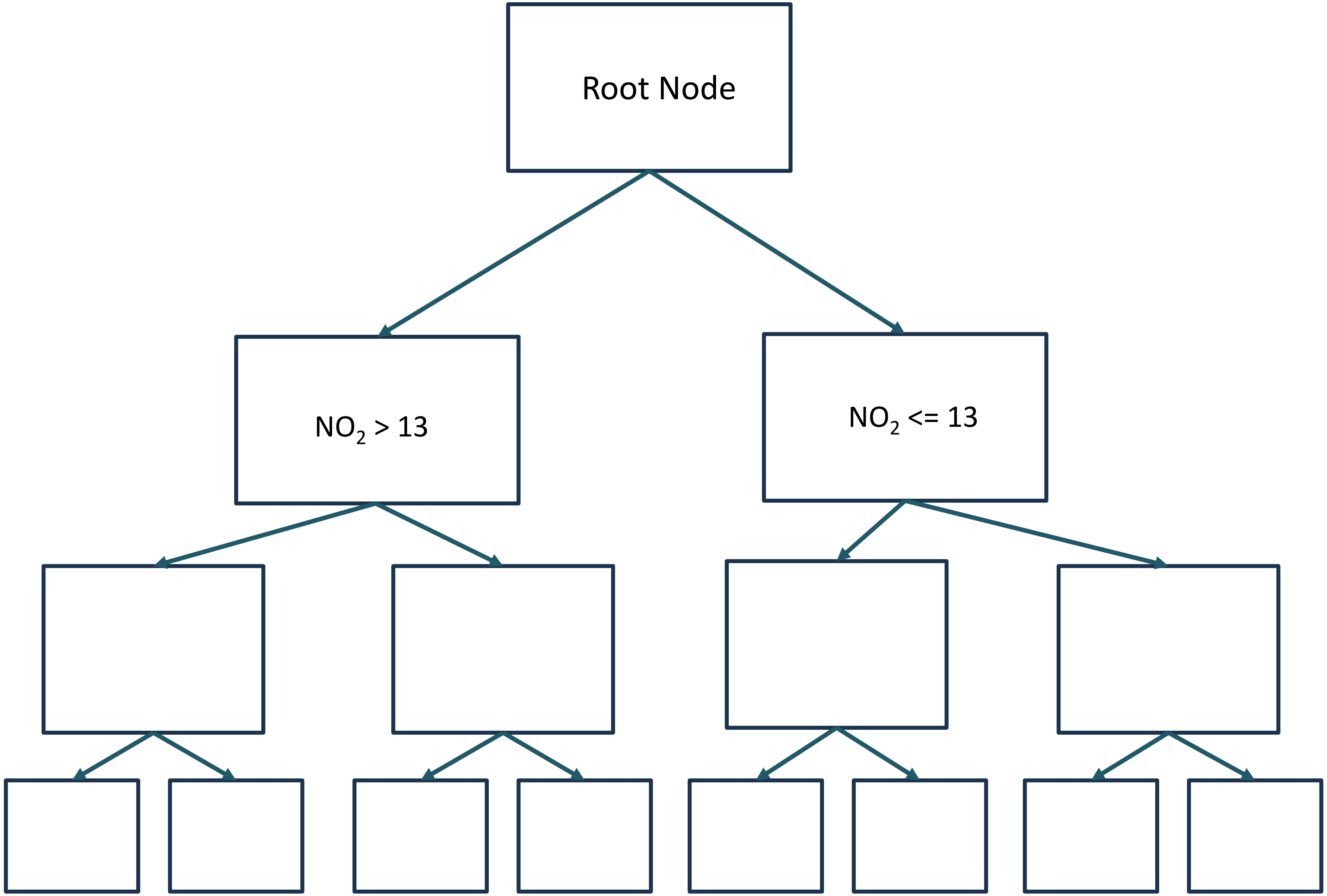
$$b = b - \left(\alpha \times \frac{\partial C}{\partial b}\right)$$

Bagging and Boosting



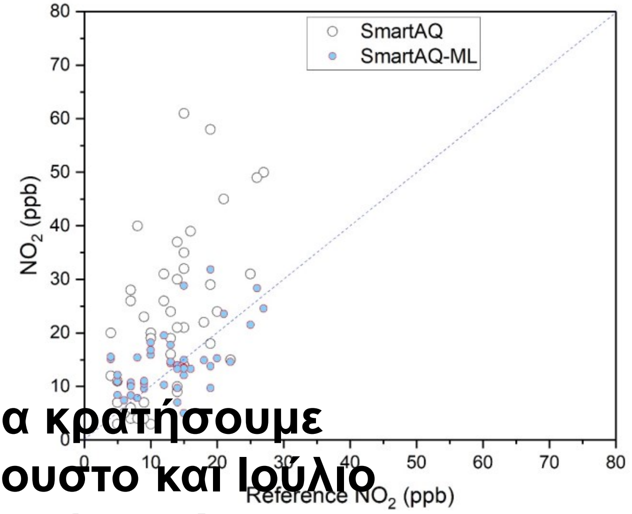


Abstraction of a tree

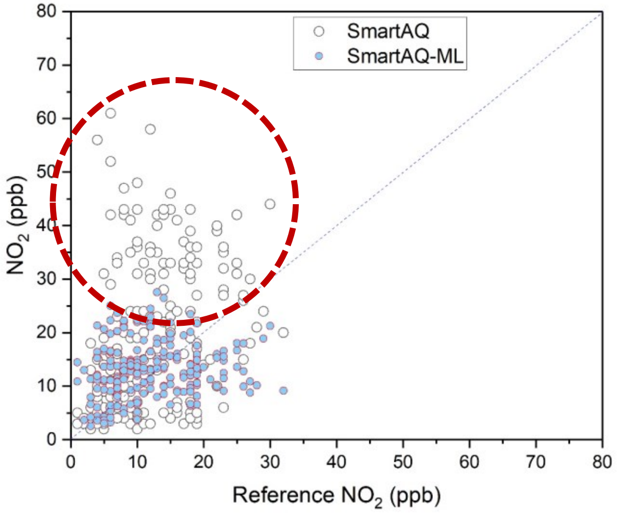


Scatter Plots

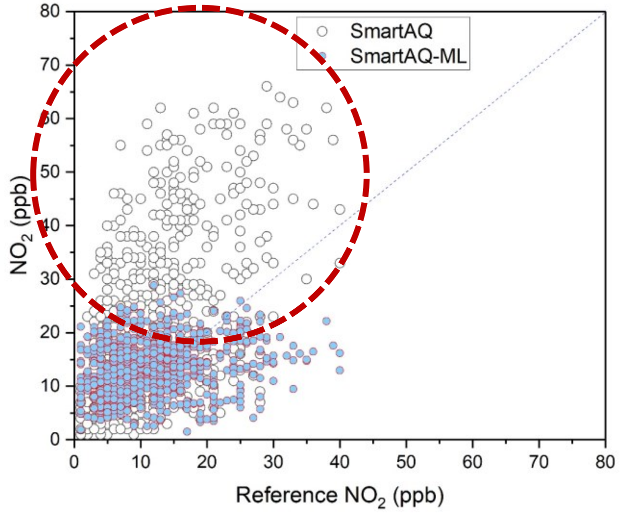
May 23



June 23

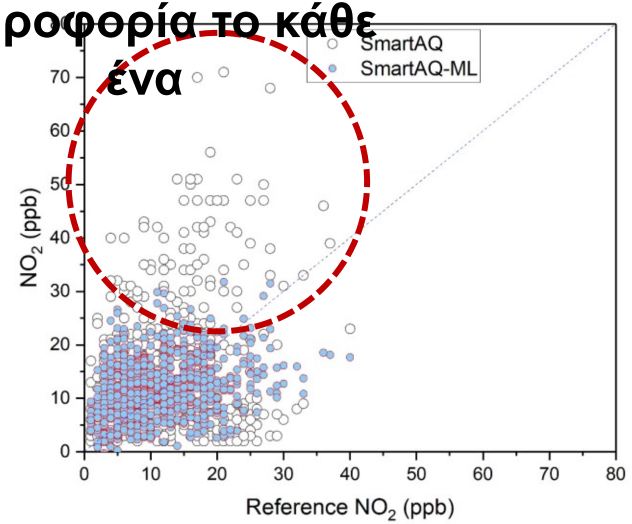


July 23

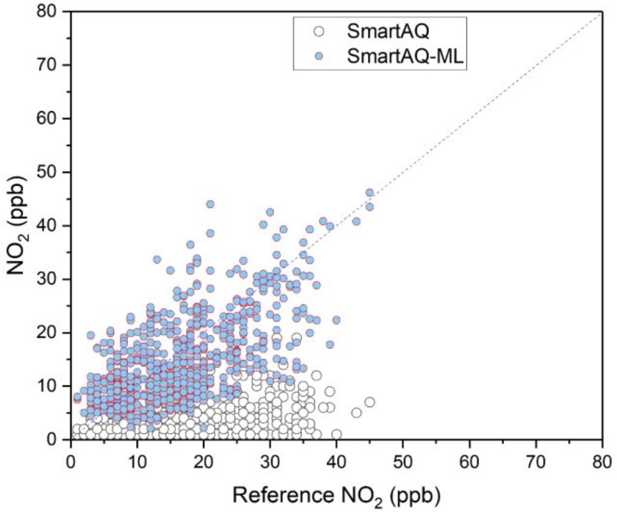


Να κρατήσουμε Αυγουστο και Ιούλιο και να ξαναγίνουν τα διαγράμματα με μία πληροφορία το κάθε ένα

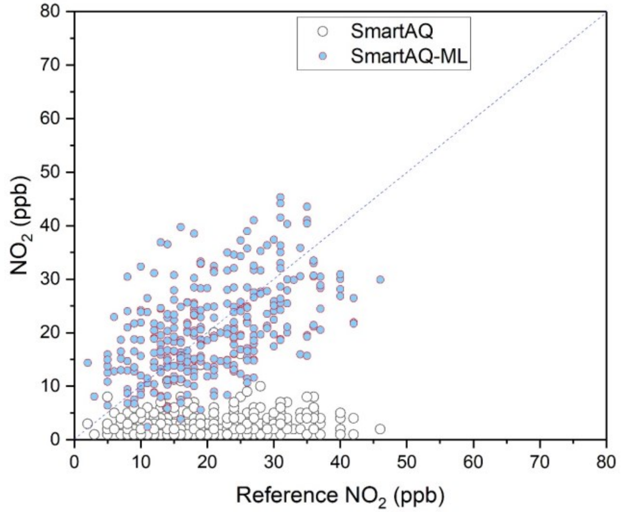
Aug 23



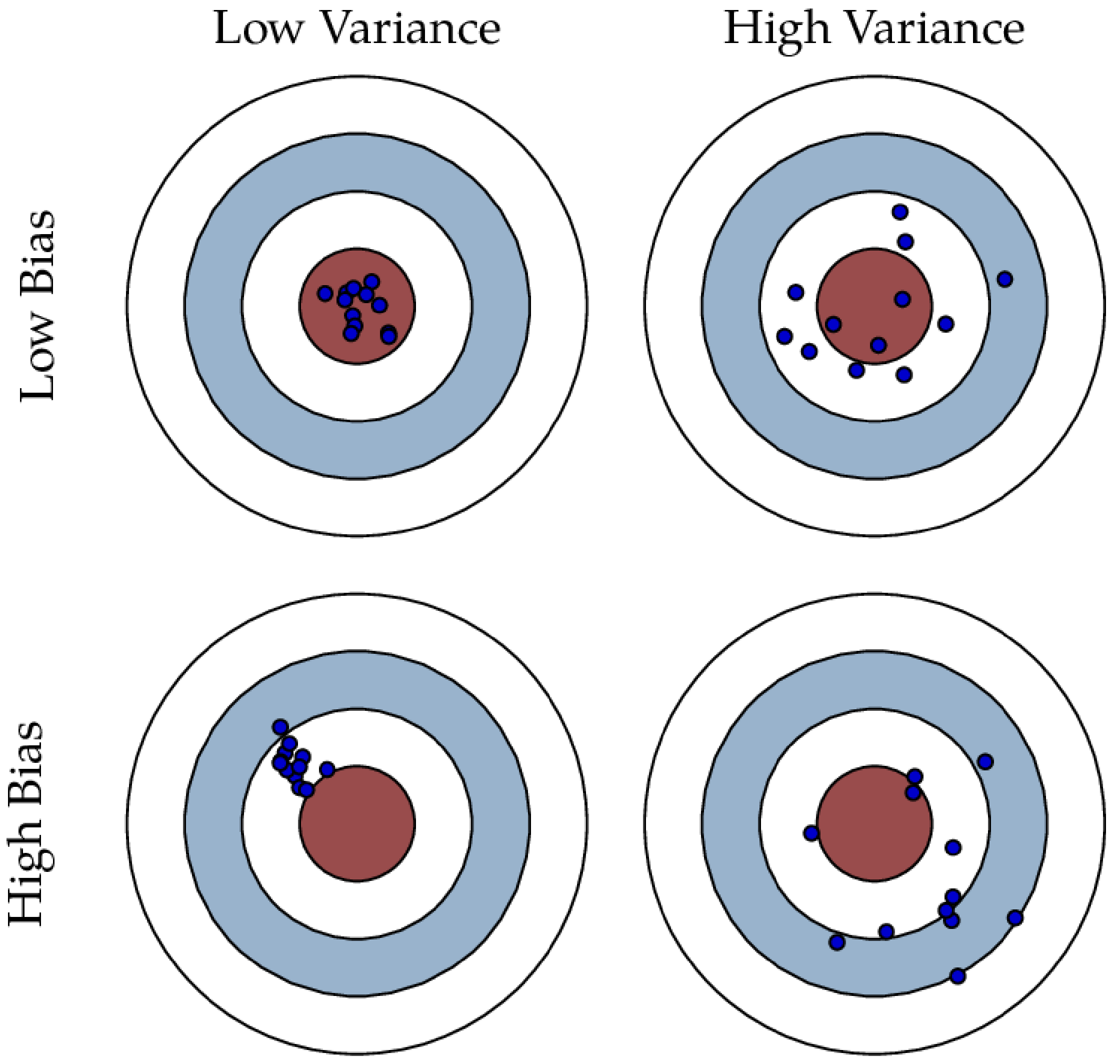
Sep 23



Oct 23

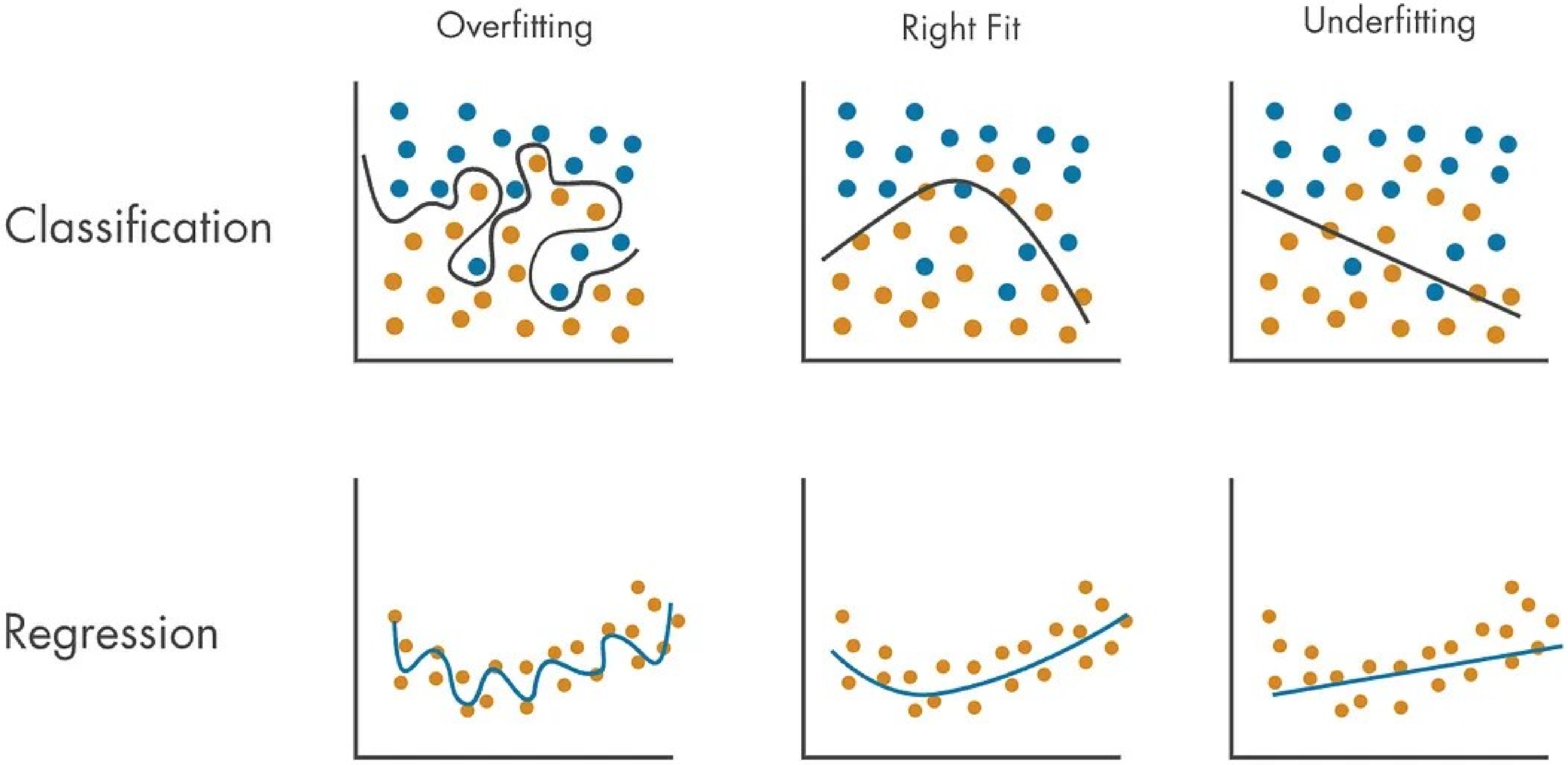


Variance and Bias



Source: endtoend.ai

Overfitting, Underfitting

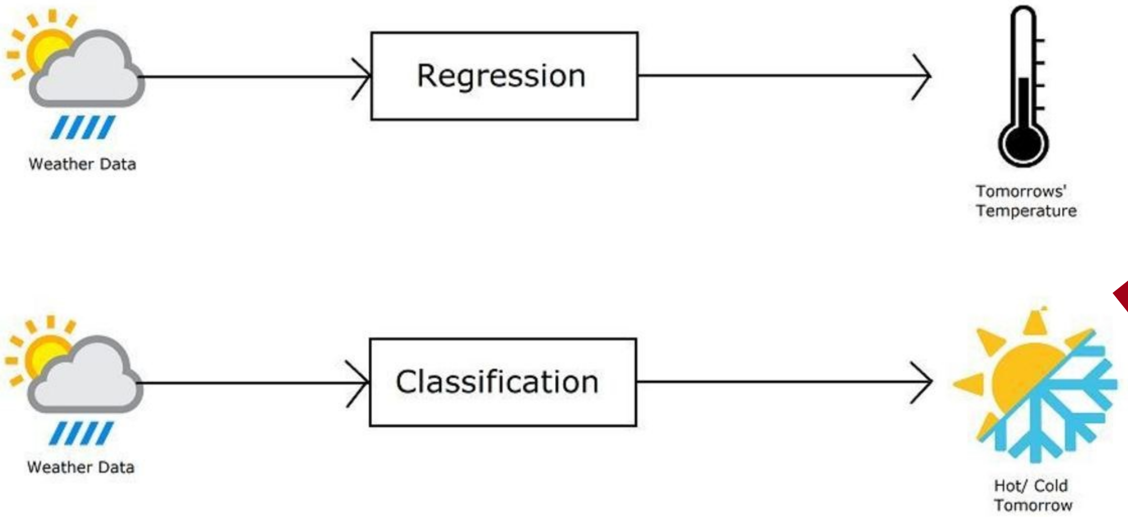


Source: MATLAB Simulink

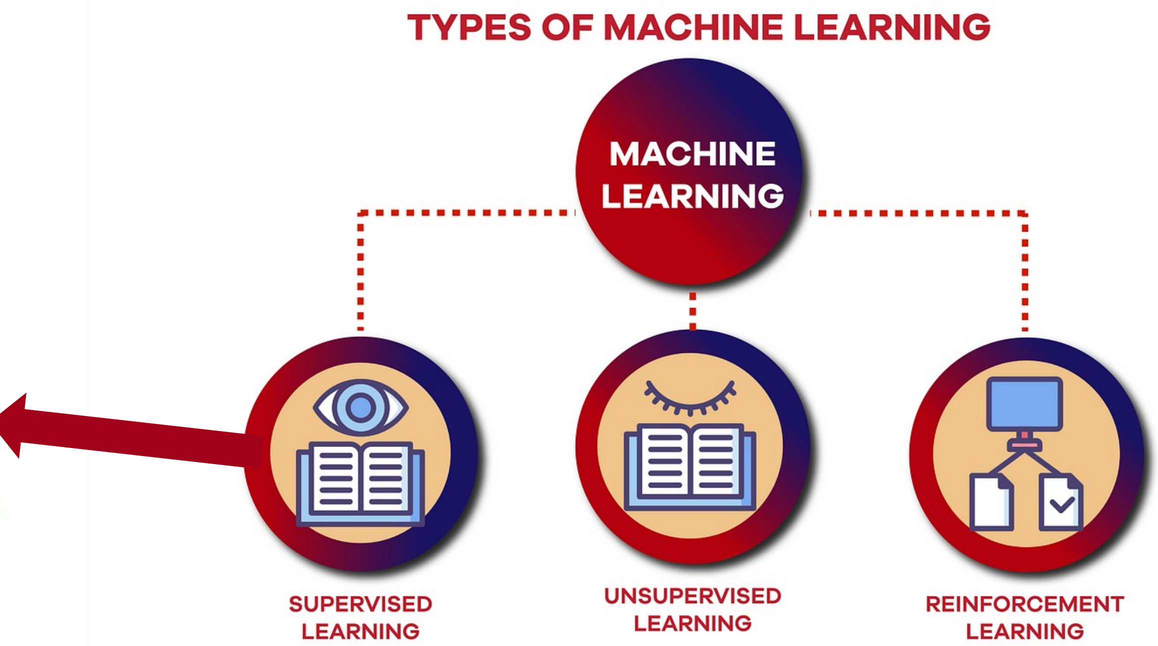
Milestones

- 1950s-1960s: The Birth of AI and Early Concepts
- 1970s-1980s: "AI Winter" and Rule-Based Systems
- 1990s: Emergence of ML Algorithms
- Late 1990s-2000s: Big Data and Boost in ML
- 2010s: Deep Learning and Neural Networks Dominate
- 2020s: Continued Advancements and Ethical Considerations

Machine Learning types



Source: turbofuture.com



Source: medium.com

Evaluation Metrics

$$ME = \frac{\sum_{i=1}^{n} |P_i - Oi|}{n} \quad (1)$$

$$MB = \frac{\sum_{i=1}^{n} (P_i - Oi)}{n} \quad (2)$$

$$FBIAS = \frac{2}{n} \sum_{i=1}^{n} \frac{(P_i - Oi)}{(P_i + O_i)} \quad (3)$$

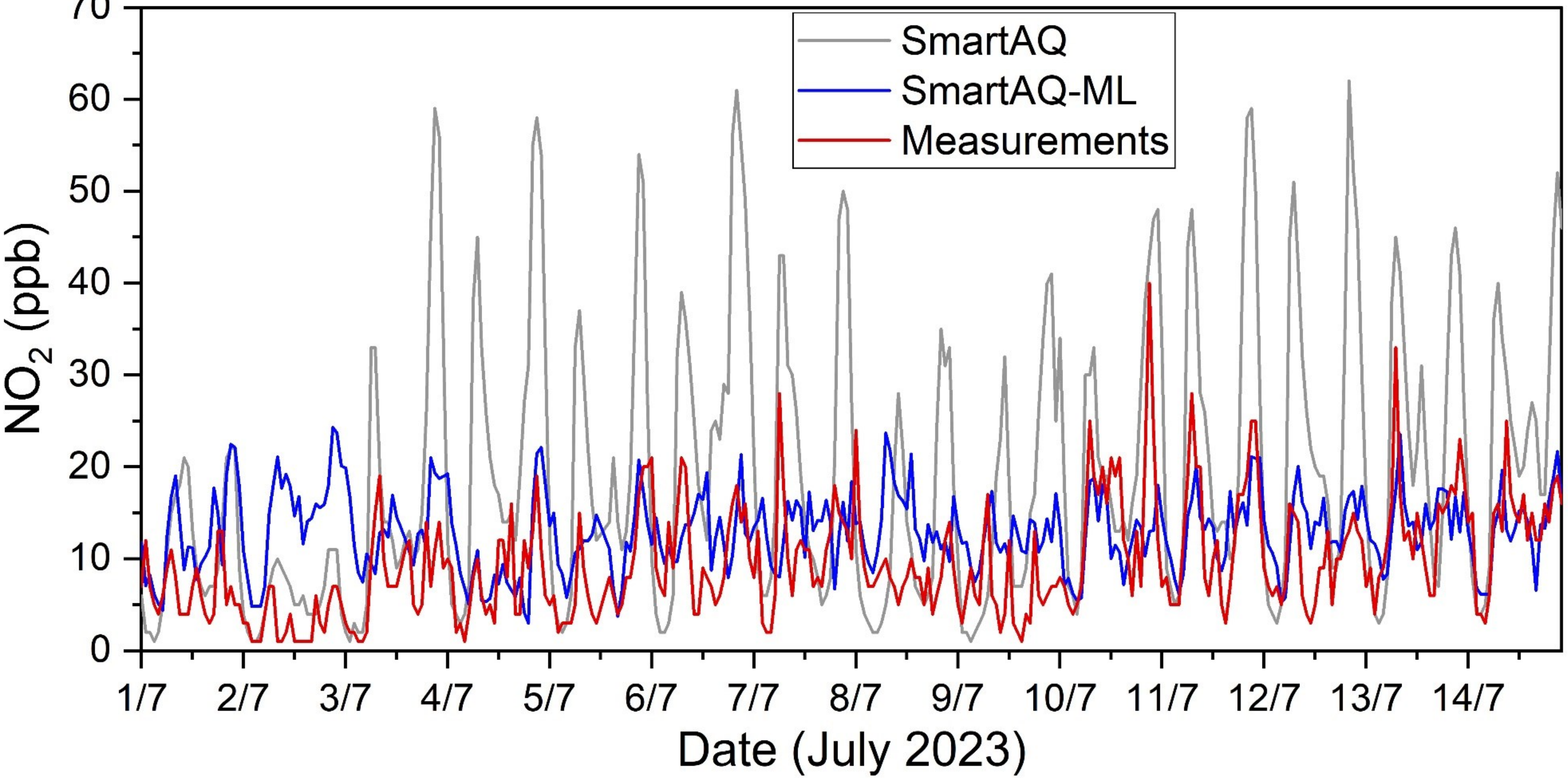
$$FERROR = \frac{2}{n} \sum_{i=1}^{n} \frac{|P_i - Oi|}{(P_i + O_i)} \quad (4)$$

$$MNE = \frac{1}{n} \sum_{i=1}^{n} \frac{|P_i - Oi|}{O_i} \quad (5)$$

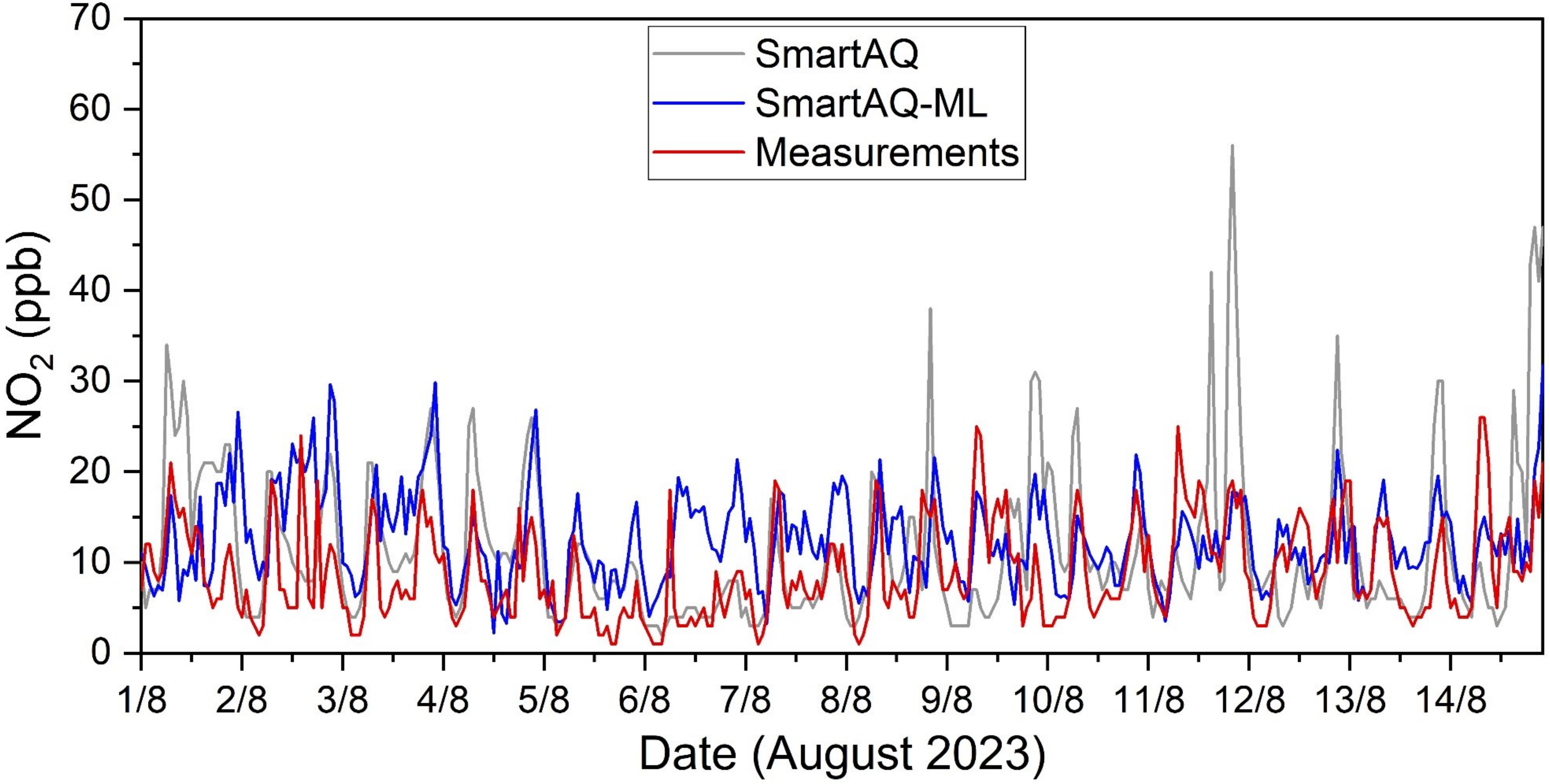
Some successful ML algorithms

- Naive Bayes
- k-Nearest Neighbors (k-NN)
- Logistic Regression
- Decision Tree
- Support Vector Machines (SVM)
- Neural Networks (Deep Learning)
- Linear Discriminant Analysis (LDA)
- Ensemble Methods (Random Forest, Gradient Boosting (e.g., XGBoost, LightGBM)

July 2023 timeseries



August 2023 timeseries



stoiximan