Forecasting total phosphorus in wastewater treatment plant effluent with machine learning – a case study in Ontario

Michael De Santi, Ciprian Panfilie, Dale Barker, Usman T. Khan,
Satinder Brar, Stephanie Gora

Collaboration

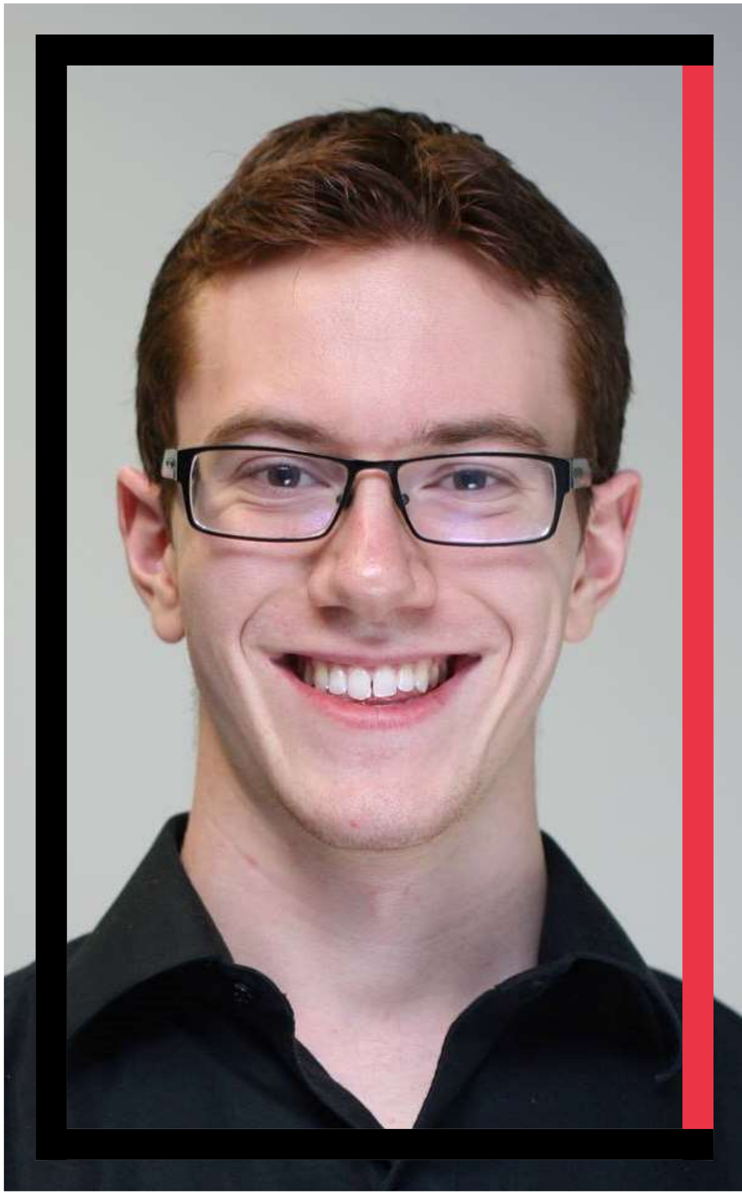
OCWA
ONTARIO CLEAN WATER AGENCY
AGENCE ONTARIENNE DES EAUX

Mitacs

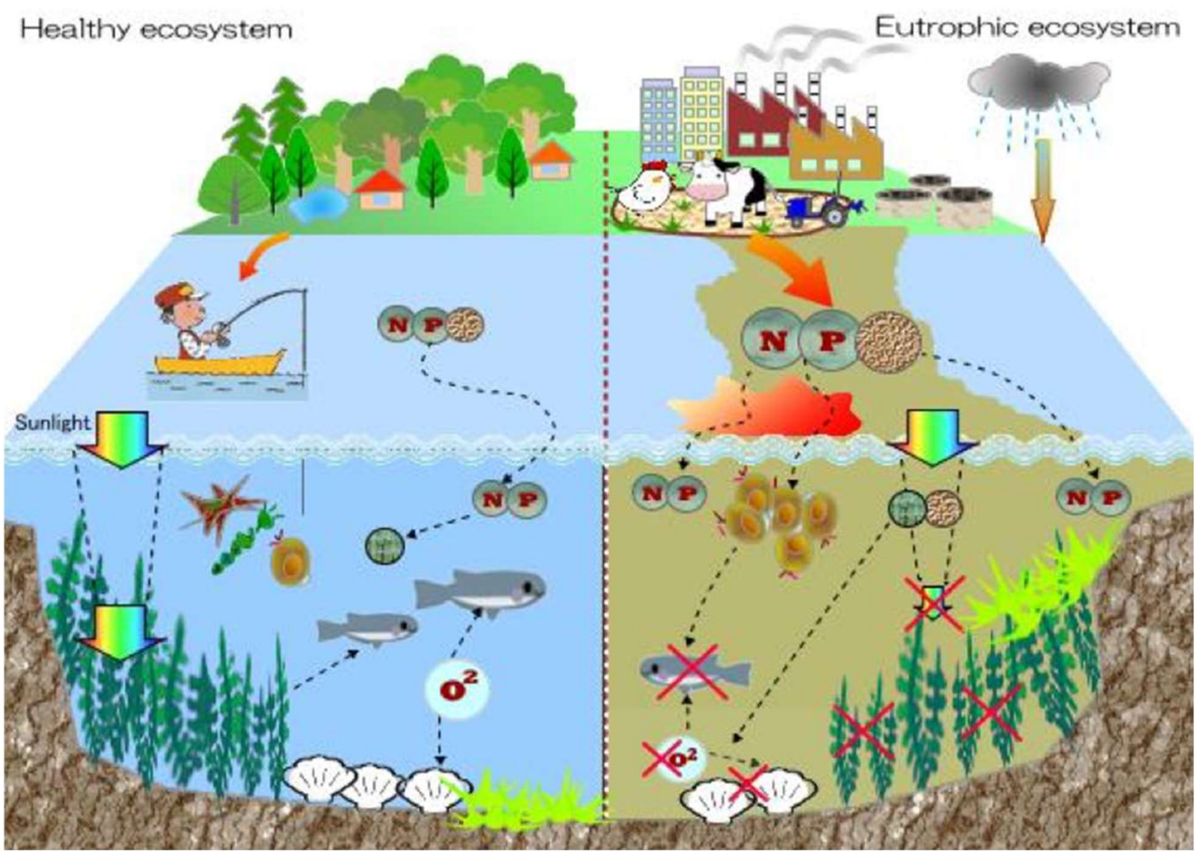
LASSONDE
SCHOOL OF ENGINEERING

YORK U





Phosphorus in the environment



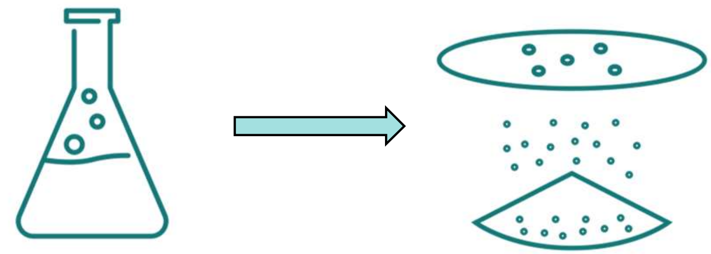
- Phosphorus in wastewater effluent has been linked to eutrophication and the development of cyanobacterial blooms
- Ontario provincial regulation: Limit total phosphorus discharge to a monthly average of *0.5 mg/L* based on composite samples taken every two weeks
- Effluent objective at WWTP: *0.3 mg/L*

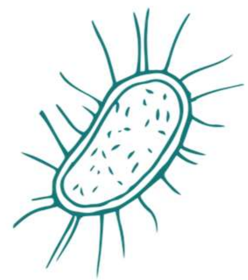
Image source: https://www.unep.org/nowpap/what-we-do/prevent-and-reduce-pollution/eutrophication

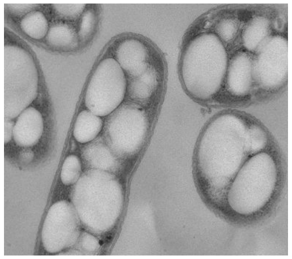
Phosphorus removal in wastewater treatment plants

Phosphorus is removed in wastewater treatment plants (WWTPs) via:

- Chemical coagulation followed by filtration
  - Lime, alum, ferric chloride
- Biological phosphorus removal
  - Enhanced biological phosphorus removal (EBPR) with phosphorus accumulating organisms (PAOs)
  - Incidental removal in secondary treatment process





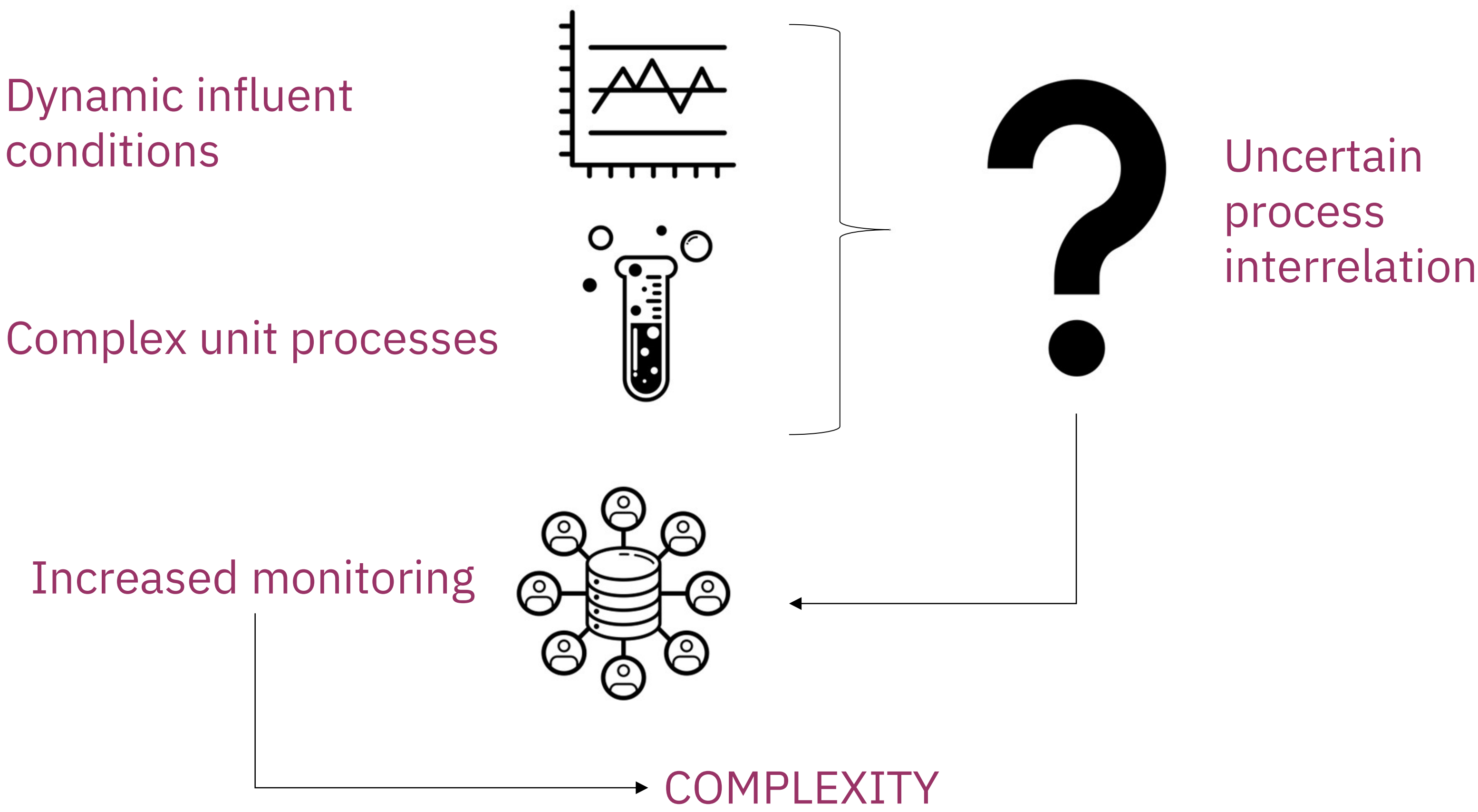


The WWTP in this study relies primarily on chemical coagulation with alum and pH adjustment at the headworks followed by filtration in the tertiary step for phosphorus removal

TEM image of PAOs used in EBPR : Yu, J., Porter, M., & Jaremko, M. (2013). Generation and utilization of microbial biomass hydrolysates in recovery and production of poly (3-hydroxybutyrate).Biomass Now: Cultivation and Utilization, InTech, 33-48.

How can we leverage existing WWTP data sources to predict effluent TP?

Wastewater treatment plants are complex



Research objectives

Objective 1: To identify the most useful process variables for predicting effluent total phosphorus (TP)

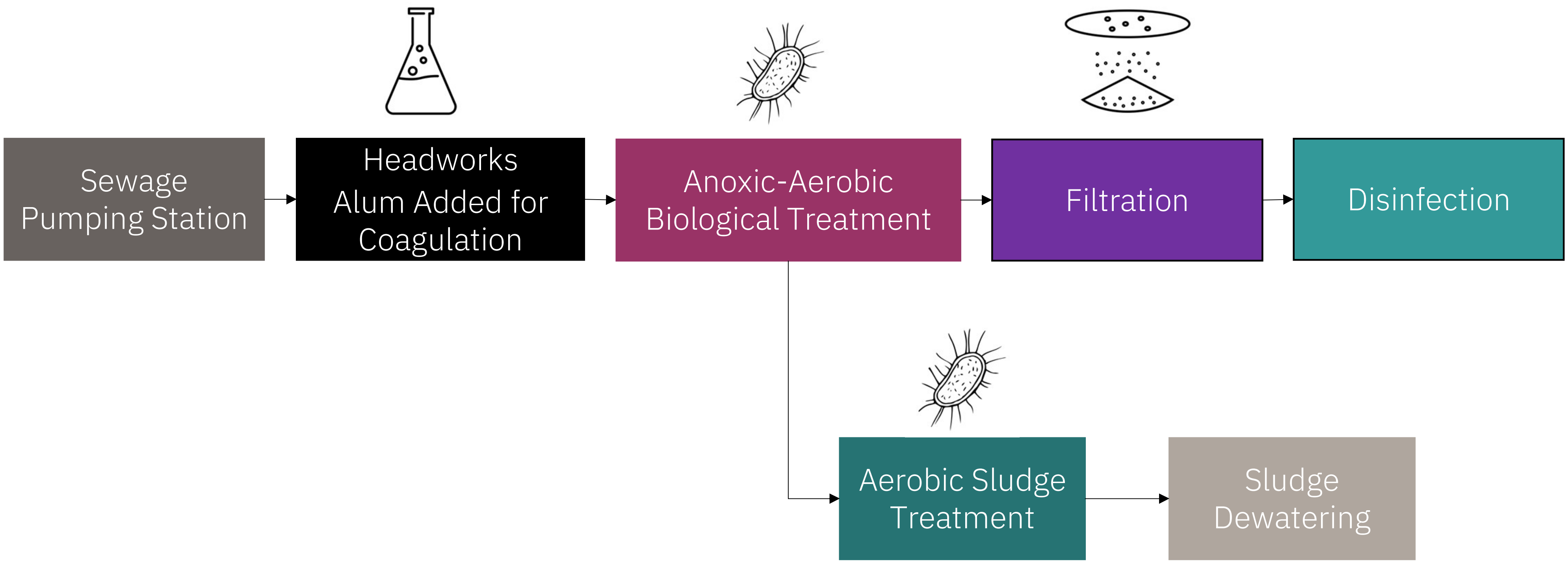
- Exploratory data analysis: Understand inputs and outputs and potential relationships
- Input variable selection: Formal process of establishing which parameters are important inputs to ML models

Objective 2: To develop and compare two approaches to modelling effluent TP using machine learning models:

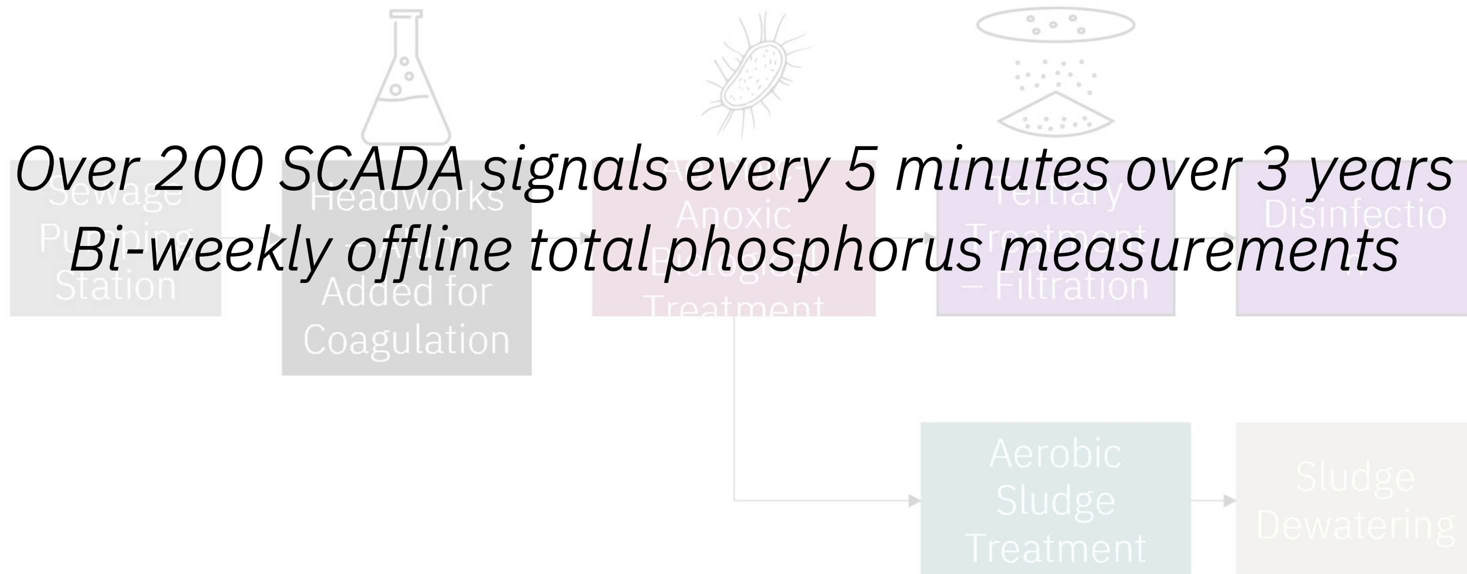
- Model 1: Predict effluent TP concentration directly
- Model 2: Predict the probability of *exceeding* the provincial effluent TP *objective*
- Model 3: Predict the probability of *exceeding* the provincial effluent TP *limit*

LASSONDE SCHOOL OF ENGINEERING | YORK U

Methods: Plant and data description



Methods: Plant and data description



Over 200 SCADA signals every 5 minutes over 3 years

Bi-weekly offline total phosphorus measurements

Modelling approach

Problem:

- What signals should we pay attention to?
- What lag period should be applied to different signals?
- Additional challenge – major discrepancy between number of SCADA measurements and number of TP measurements:
  - SCADA: once every 5 minutes = 105,120 measurements per year
  - TP: twice weekly measurements = 104 measurements per year
  - Ratio is 1000:1!

Modelling approach

Solution:

- Exploratory data analysis
- Prepare the data
  - Average SCADA inputs over the course of the day
- Build models
  - Regression
  - Classification
- Input variable selection (IVS)
  - Implement multi-method approach considering model-based and model-free methods with a final iterative backwards elimination approach to identify important input variables

Iterative

Preliminary data exploration

We reviewed:

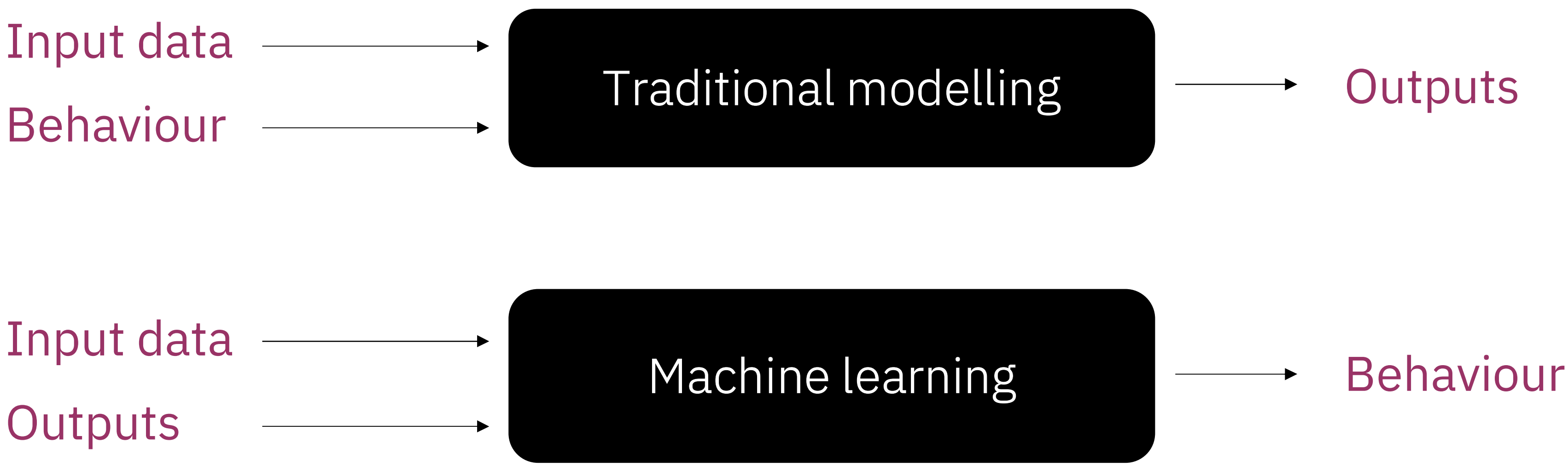
- Influent water quality
- Effluent water quality
- SCADA inputs/outputs
- Potential relationships between these

Solution: Machine learning

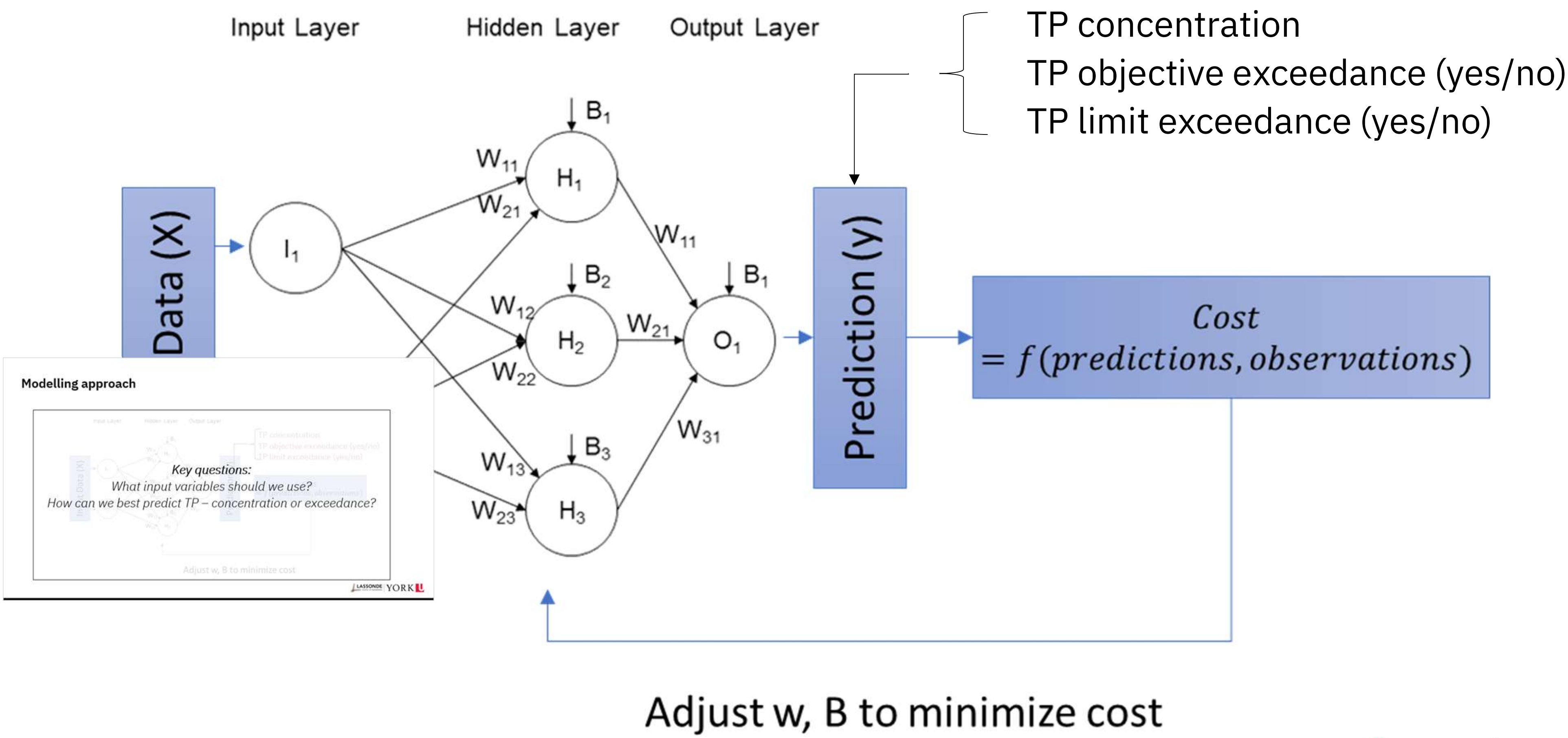
Machine learning:

"The task of showing the inputs and outputs of a problem to an algorithm and letting it learn how to solve it"

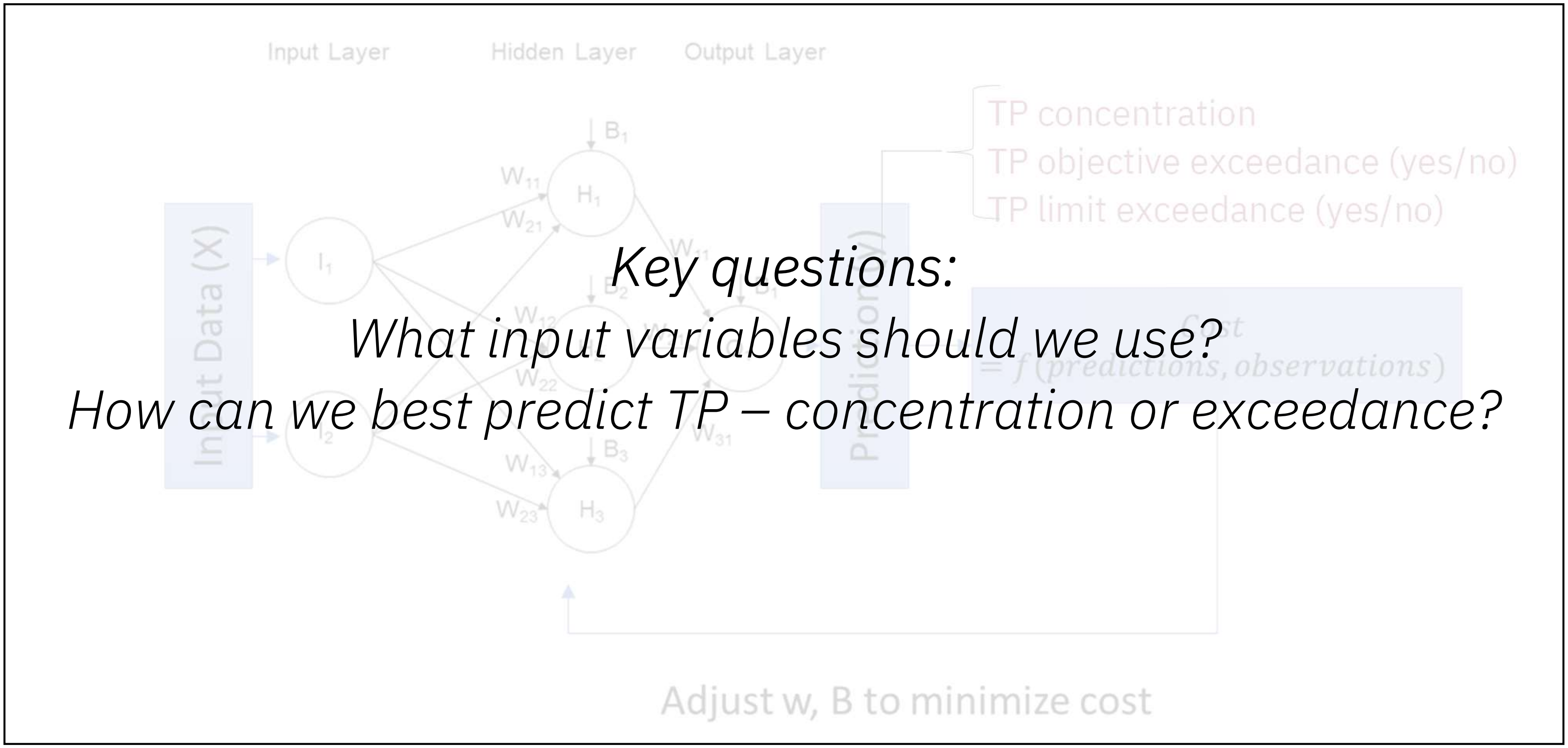
- Serpa (2020) in *Towards Data Science*



Modelling approach



Modelling approach



Key questions:

What input variables should we use?

How can we best predict TP – concentration or exceedance?

Build initial models

Each model included multiple artificial neural network (ANN) multilayer perceptron (MLP) *base learners* that were combined to create *ensemble* ANNs

For each ensemble model, the response from each base learner is combined to create a confidence interval or probabilistic output

Model 1 Predict TP concentration

- Regression model
- Cost function = mean squared error

Model 2 Predict exceedance of TP objective

- Classification model
- Cost function = binary cross entropy

Model 3 Predict exceedance of TP reg.

- Classification model
- Cost function = binary cross entropy

Evaluating model performance

Regression models*:* Root mean square error (RMSE) and R^2

$$\mathbf{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\mathrm{Predicted}_i - \mathrm{Observed}_i)^2}{n}}$$

Classification models: Accuracy, Recall, Precision, Brier Score

$$\mathbf{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{n}$$

$$\mathbf{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$\mathbf{Precision} = \frac{\text{True Positives} + \text{False Positives}}{\text{True Positives}}$$

$$\mathbf{Brier\ score} = \sum_{i=1}^{n} \frac{\left(y_i^{obs} - p_i\right)^2}{n}$$

Confusion matrix

Predicted Condition	
Positive	Negative
True Positive	False Negative
False Positive	True Negative

Summary of model results

Model	Type	Performance	Notes
Model 1: TP concentration	Regression	RMSE: 0.299 mg/L	Mean error was about 0.3 mg/L +/- the predicted value – this is pretty good!
Model 2: Exceedance of TP objective	Classification	Accuracy: 71% Recall: 73% Precision: 68%	Correctly predicted exceedance vs. non-exceedance 71% of the time and identified exceedances 73% of the time
Model 3: Exceedance of TP regulatory limit	Classification	Accuracy: 71% Recall: 53% Precision: 47%	Correctly predicted exceedance vs. non-exceedance 71% of the time and identified exceedances 53% of the time

Takeaways

1. ML can accurately forecast wastewater effluent quality using routinely collected data
2. ML can select unintuitive but useful relationships between process variables and target outputs
3. The highly correlated and non-linear relationships between process variables require advanced IVS

Acknowledgements

OCWA

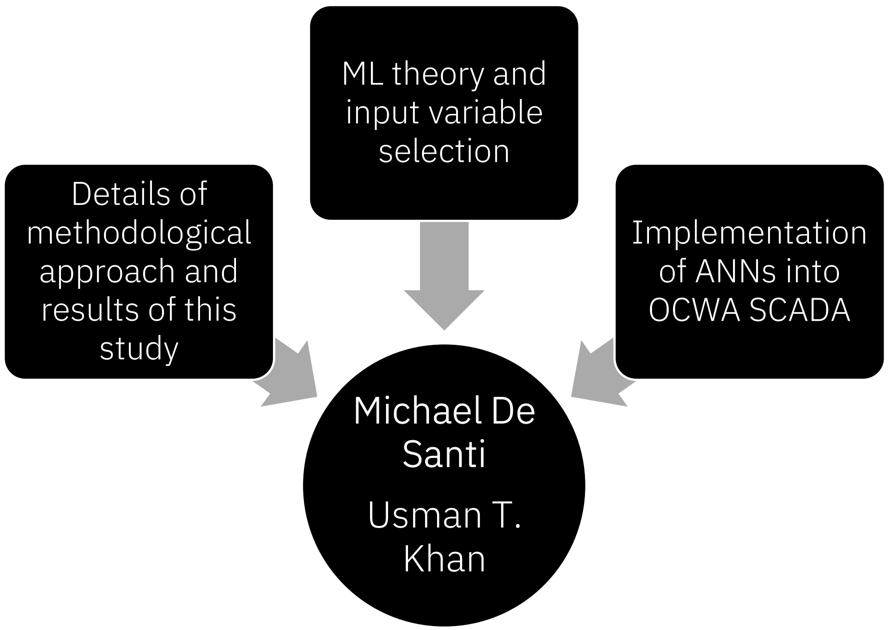
ONTARIO CLEAN WATER AGENCY

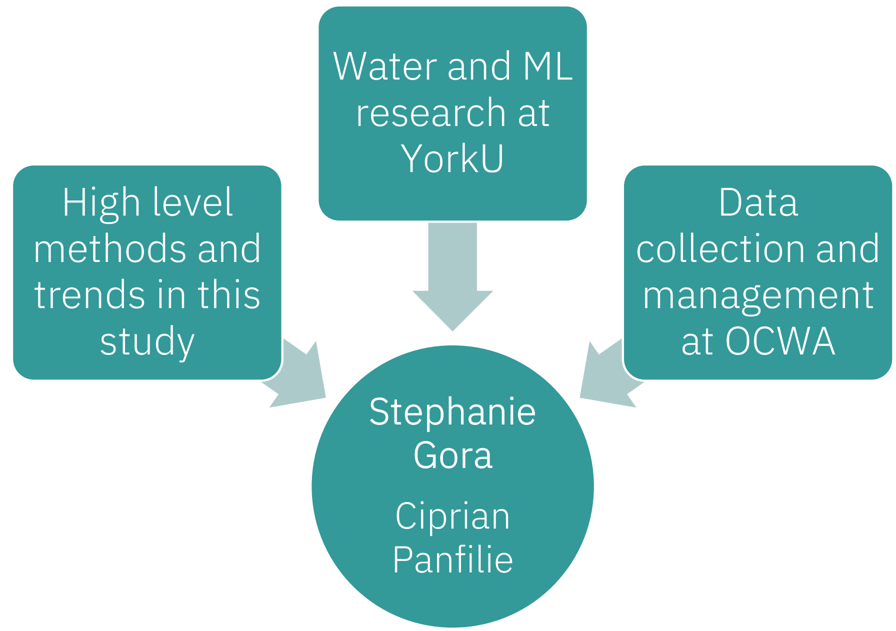
AGENCE ONTARIENNE DES EAUX

Mitacs

LASSONDE SCHOOL OF ENGINEERING | YORK U

Questions? Comments?





Email: desantim@yorku.ca
Twitter: @michaelvdesanti
LinkedIn: www.linkedin.com/in/michaeldesanti/

Email: stephanie.gora@lassonde.yorku.ca
Twitter: @goraradio
LinkedIn: www.linkedin.com/in/stephaniegora/