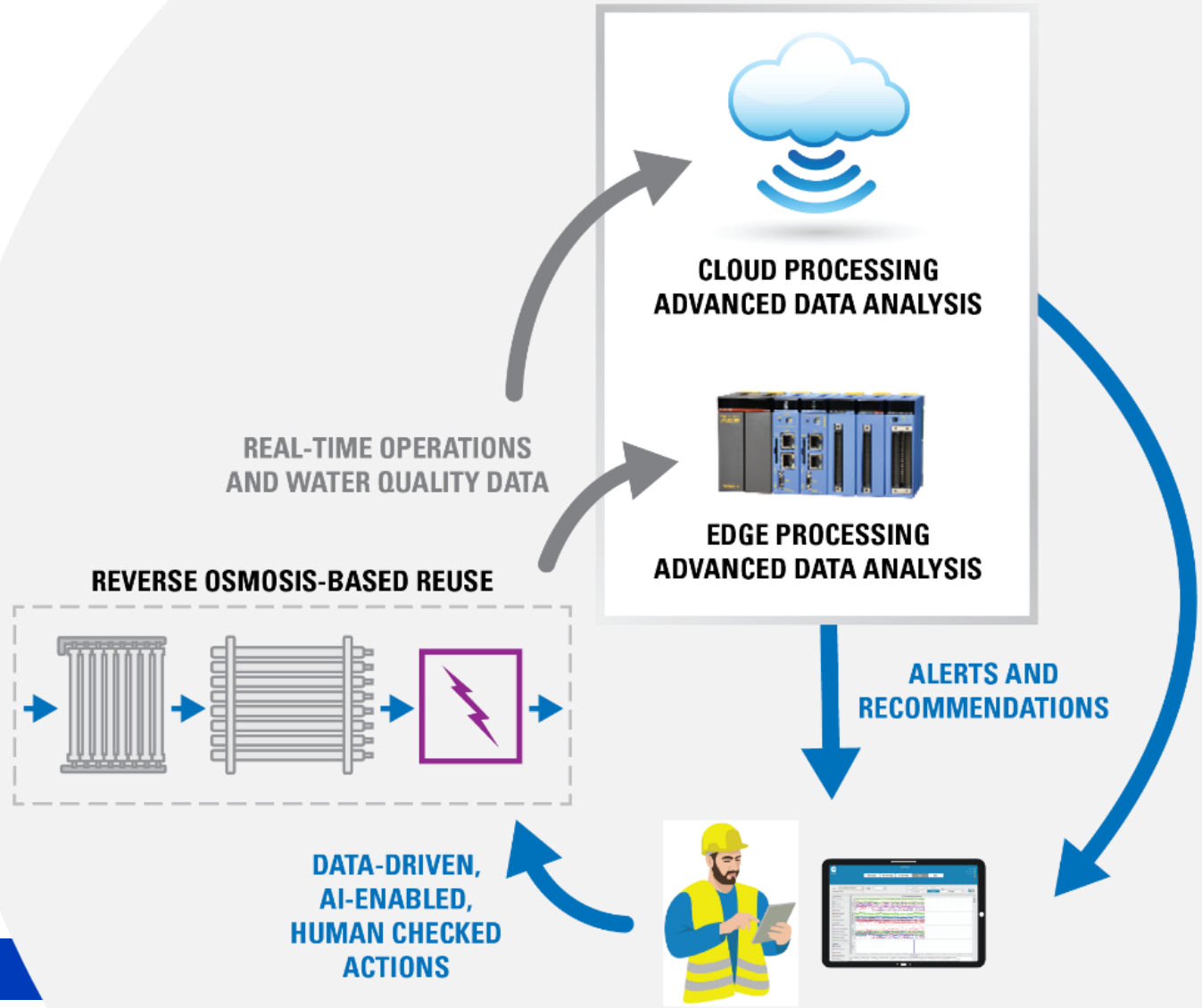
Texas AWWA Water Science
Lunch Break Seminar

Machine Learning for Potable Reuse

Kyle Thompson, PhD, PE
kthompson@carollo.com
Austin, TX

August 30, 2023



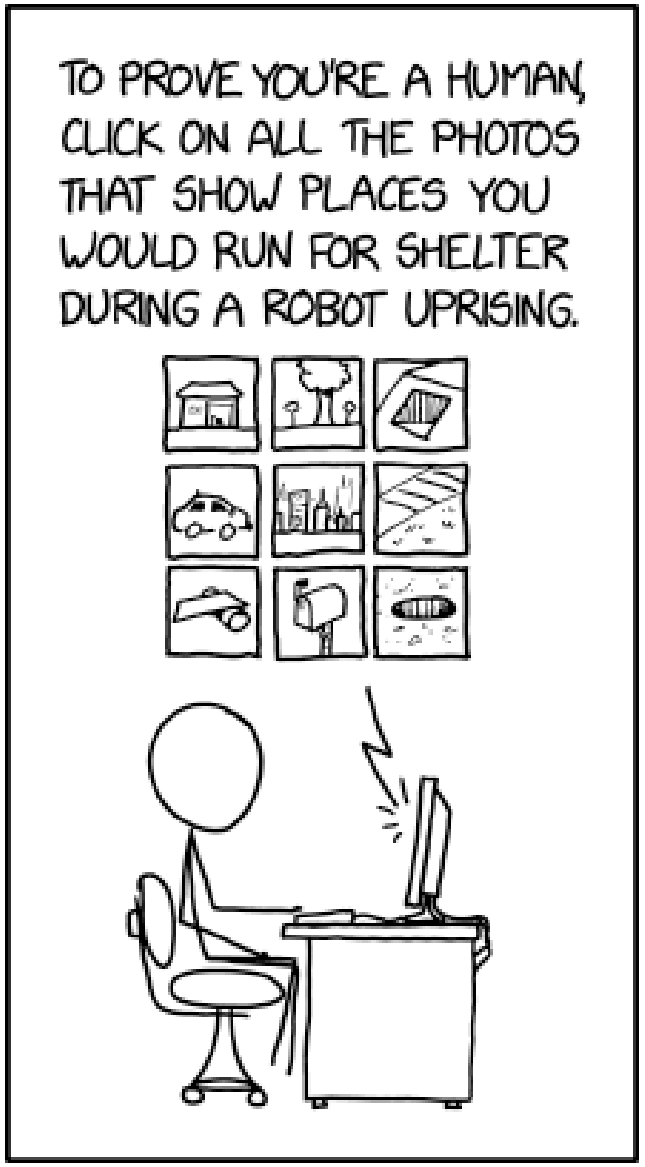
Presentation Outline

1) What is machine learning?
2) Why apply machine learning?
3) How does machine learning work?
4) What have been some uses of machine learning for reuse?
5) How should we go about applying machine learning or AI?

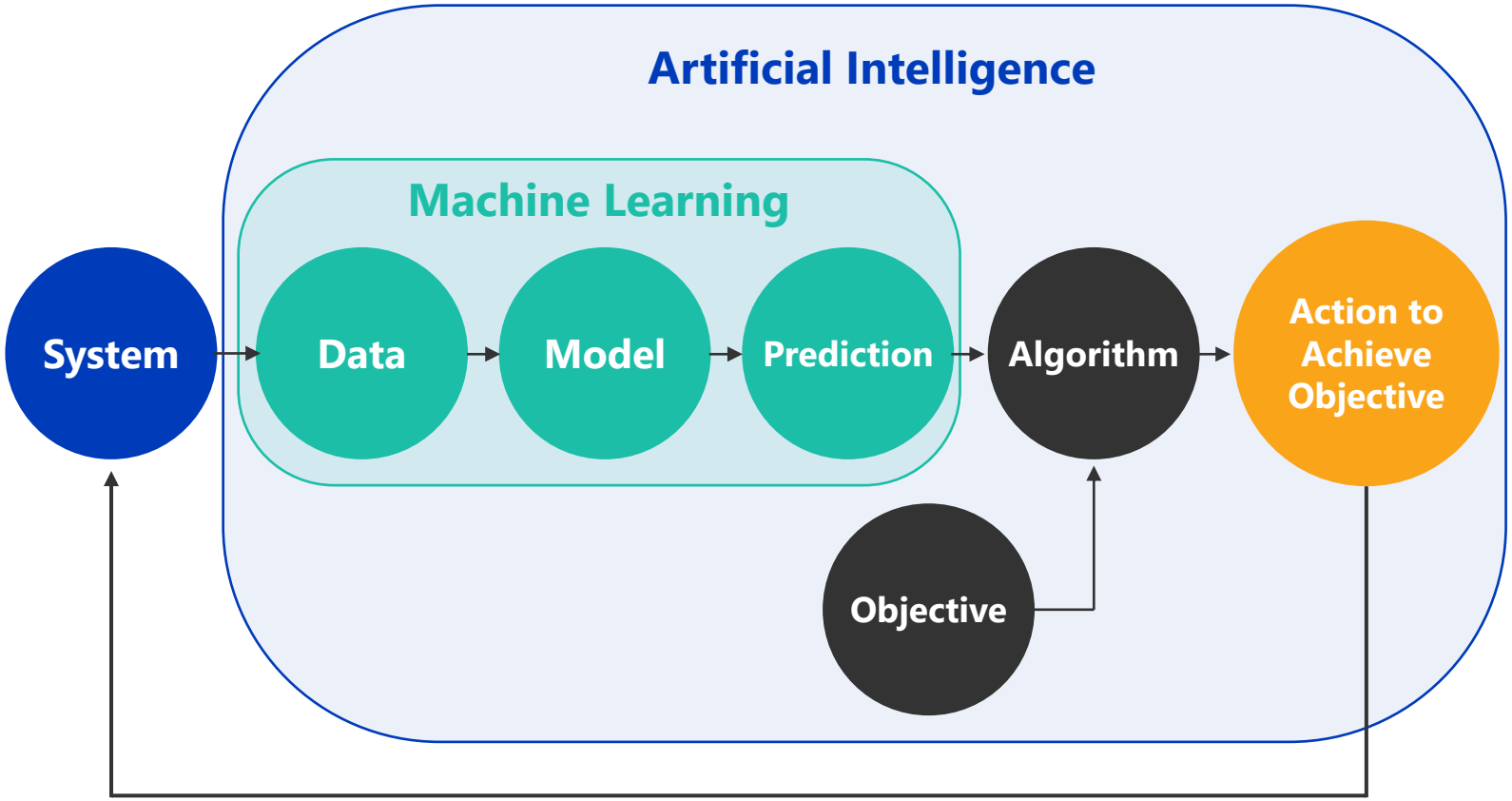
1 What is machine learning?

Machine Learning (ML) is the study and application of algorithms that learn from and make predictions based on data

- Characteristics/Benefits
  - » Data-driven
  - » Adaptive
  - » Multivariate
  - » Nonlinear

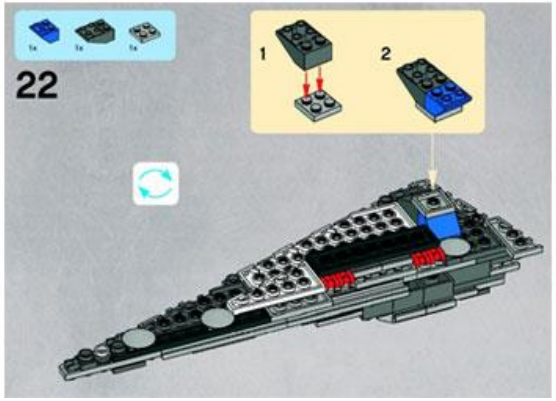


What's the difference between ML & AI?



Machine learning approaches

Supervised



Unsupervised

Two types of supervised machine learning

Classification

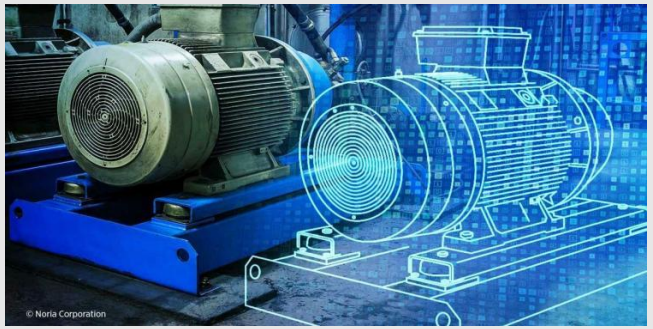
- *Categorical* outputs
- Is this a dog or a cat?
- Useful for alarms

Regression

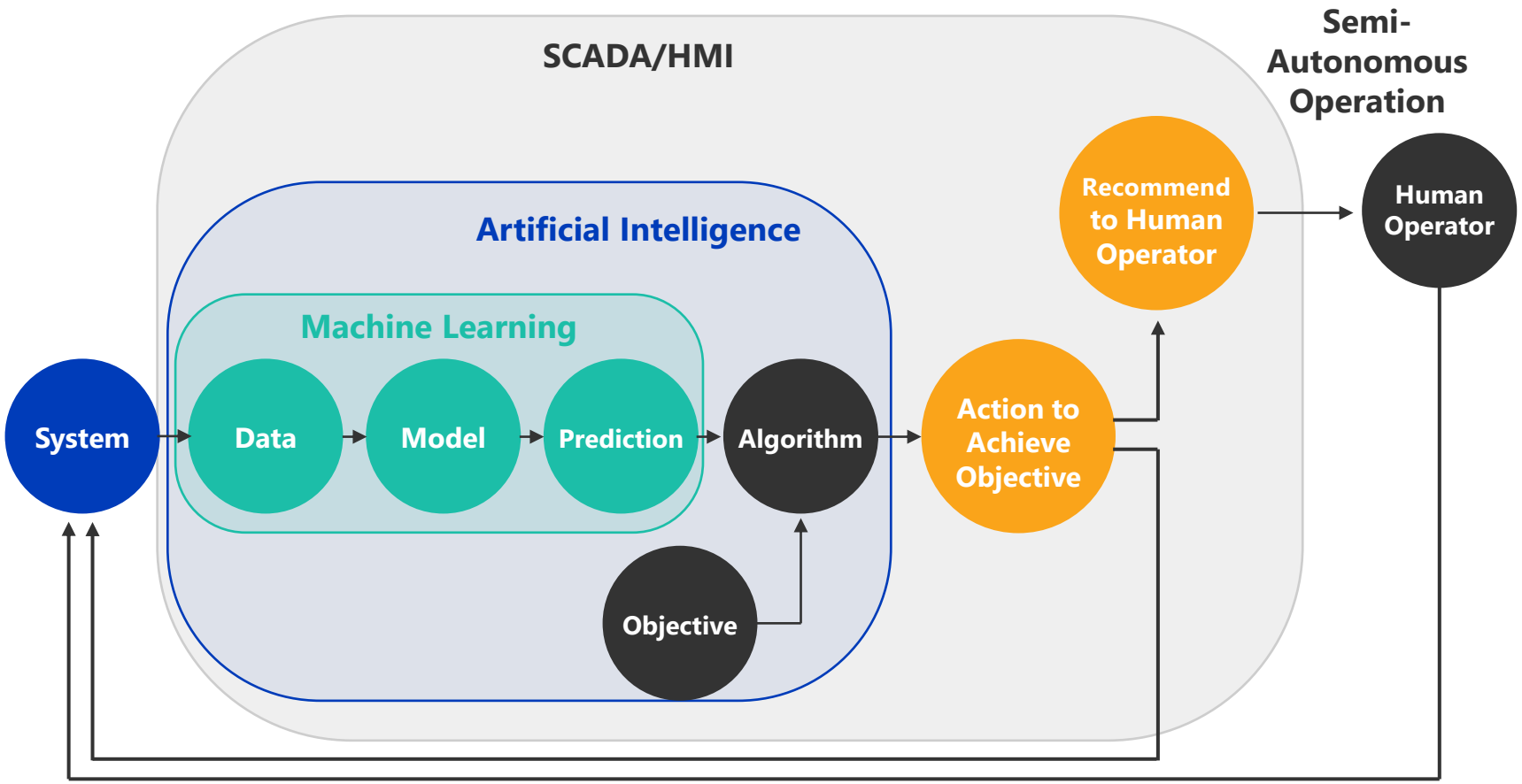
- *Numerical* outputs
- You open 2.5 cans of tuna and 1.4 bags of catnip. How many cats run to your kitchen?
- Useful for process control

Machine learning implementation styles

- **Fault Detection** – Monitoring for when an issue has occurred through pattern recognition on sensor data
- **Soft Sensors** – Predicting a slower or more expensive contaminant of concern with faster or cheaper data
- **Digital Twin** – A model with automated, bidirectional data connectivity to allow both real-time model updates and control adjustments



Semi-autonomous operation

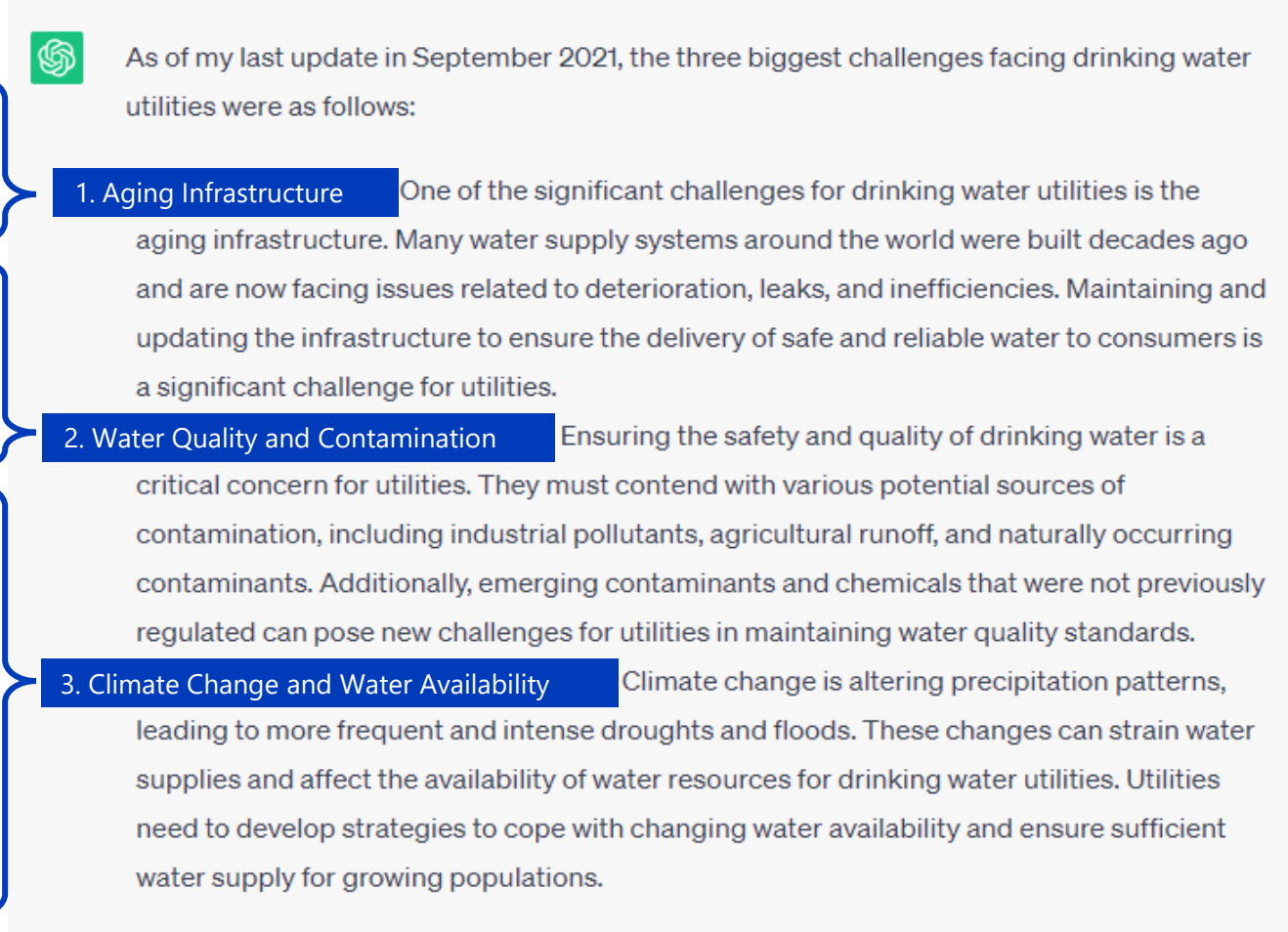


2

Why apply machine learning?

What are the three biggest challenges facing drinking water utilities?

- Predict location of lead service lines
- Predict pipes likely to leak
- Soft sensors to predict concentrations of contaminants
- Model GAC PFAS breakthrough
- Predict water flows
- Digital twins that respond to changes in temperature, turbidity, etc.
- **Optimize cost, energy, and reliability of advanced treatment for reuse**



As of my last update in September 2021, the three biggest challenges facing drinking water utilities were as follows:

1. Aging Infrastructure: One of the significant challenges for drinking water utilities is the aging infrastructure. Many water supply systems around the world were built decades ago and are now facing issues related to deterioration, leaks, and inefficiencies. Maintaining and updating the infrastructure to ensure the delivery of safe and reliable water to consumers is a significant challenge for utilities.

2. Water Quality and Contamination: Ensuring the safety and quality of drinking water is a critical concern for utilities. They must contend with various potential sources of contamination, including industrial pollutants, agricultural runoff, and naturally occurring contaminants. Additionally, emerging contaminants and chemicals that were not previously regulated can pose new challenges for utilities in maintaining water quality standards.

3. Climate Change and Water Availability: Climate change is altering precipitation patterns, leading to more frequent and intense droughts and floods. These changes can strain water supplies and affect the availability of water resources for drinking water utilities. Utilities need to develop strategies to cope with changing water availability and ensure sufficient water supply for growing populations.

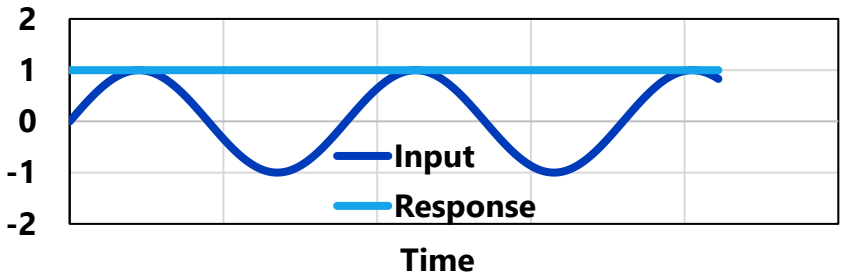
Tools like ML make water treatment more *data-driven*

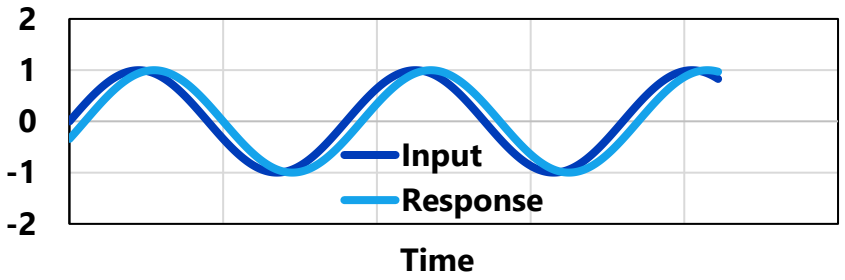
Reduce Downtime

Reactive → Predictive

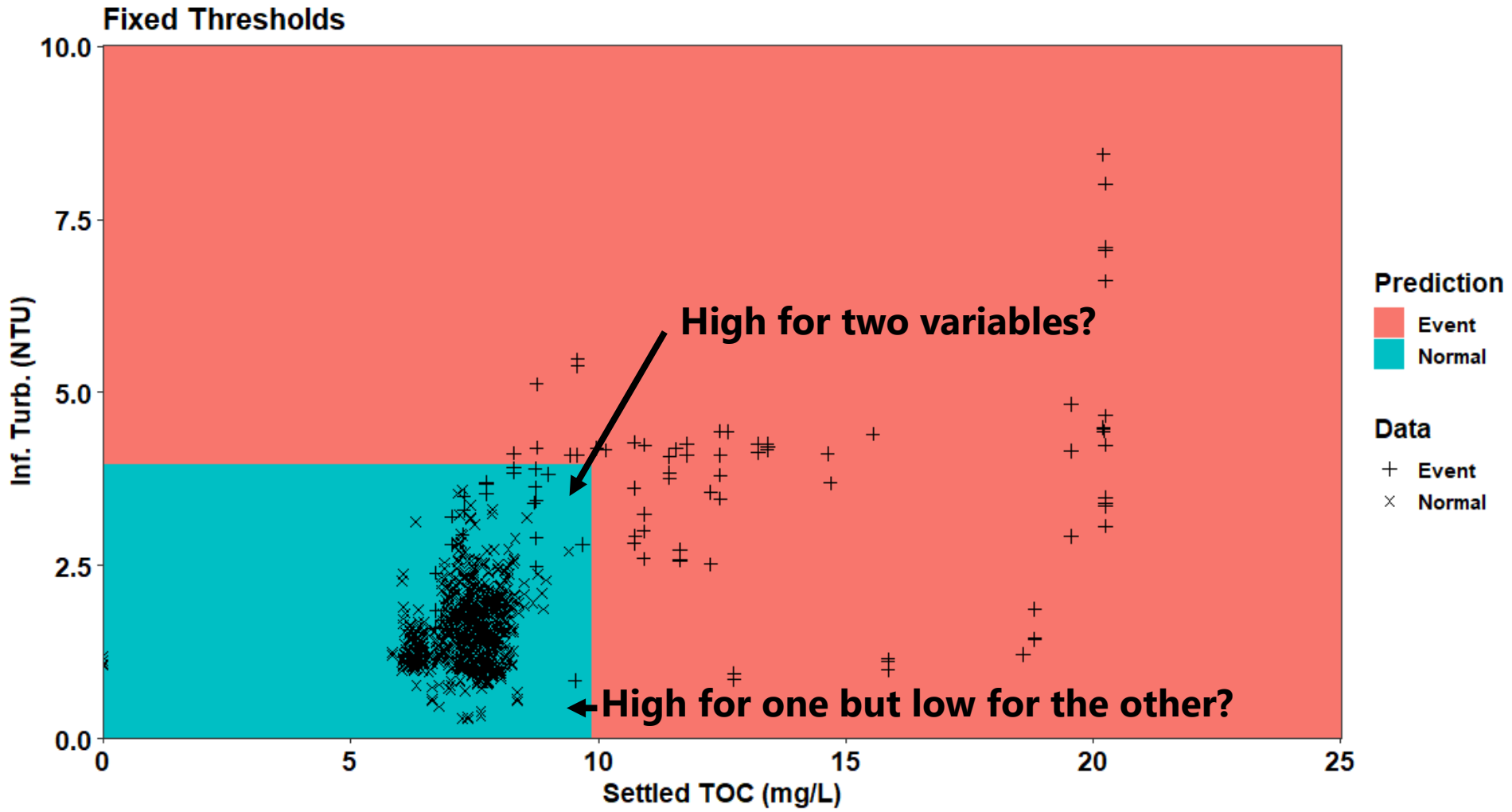
Set-It-And-Forget-It → Adaptive

Save Cost and Energy

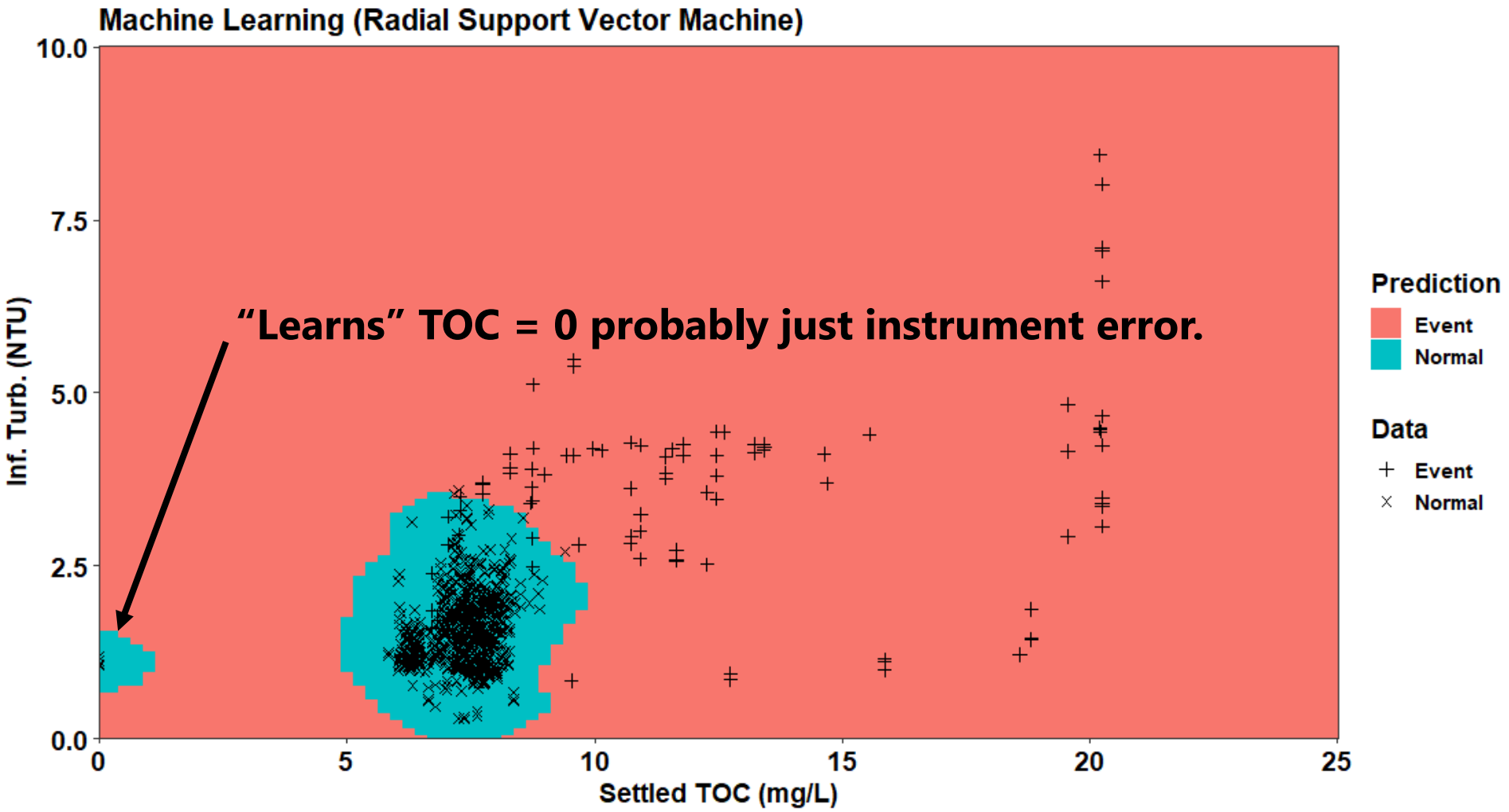




Conventional alarms set fixed thresholds on individual variables



Machine learning can construct adaptive, non-linear boundaries between normal and alarms



3

How does machine learning work?

There are lots of types of machine learning

6 Available Models

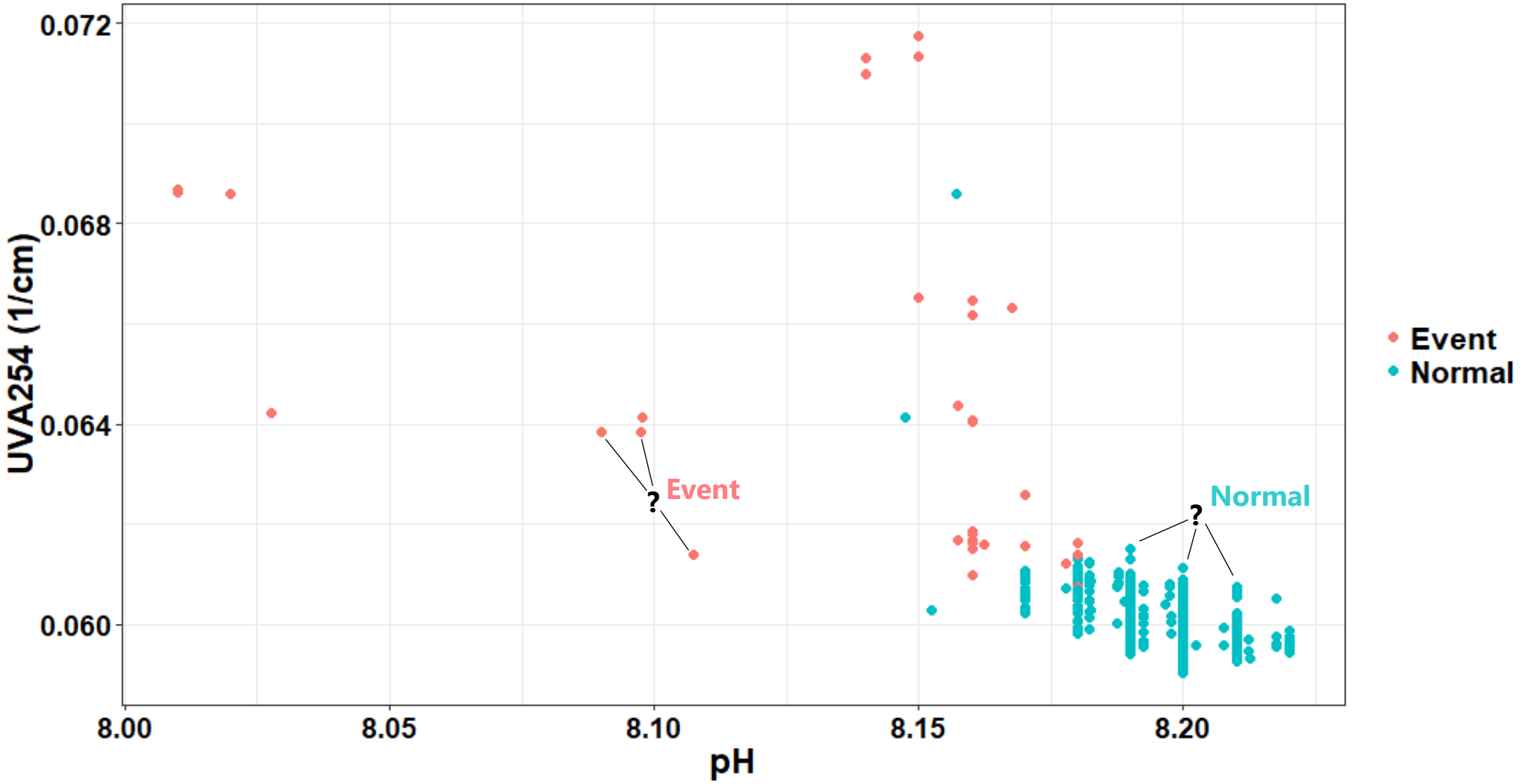
The models below are available in `train`. The code behind these protocols can be obtained using the function `getModelInfo` or by going to the github repository.

Show 238 entries

Search:

Model	*method* Value	Type	Libraries	Tuning Parameters
AdaBoost Classification Trees	adaboost	Classification	fastAdaboost	nIter, method
AdaBoost.M1	AdaBoost.M1	Classification	adabag, plyr	mfinal, maxdepth, coeflearn
Adaptive Mixture Discriminant Analysis	amdai	Classification	adaptDA	model

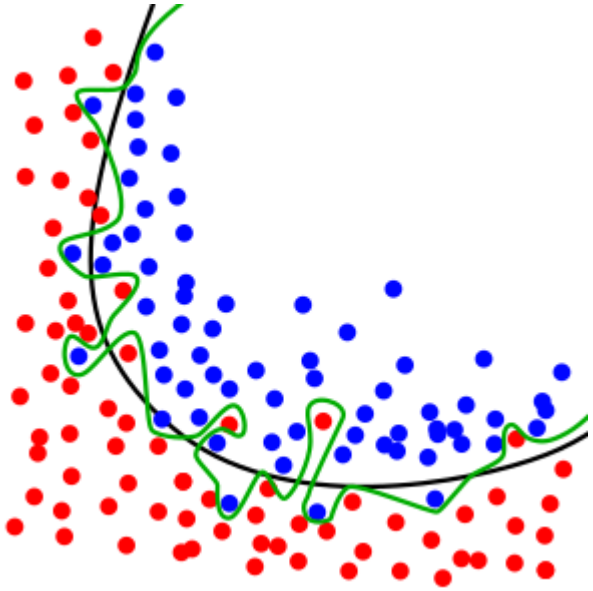
k-Nearest Neighbors guesses data is the same category as the nearest neighbors



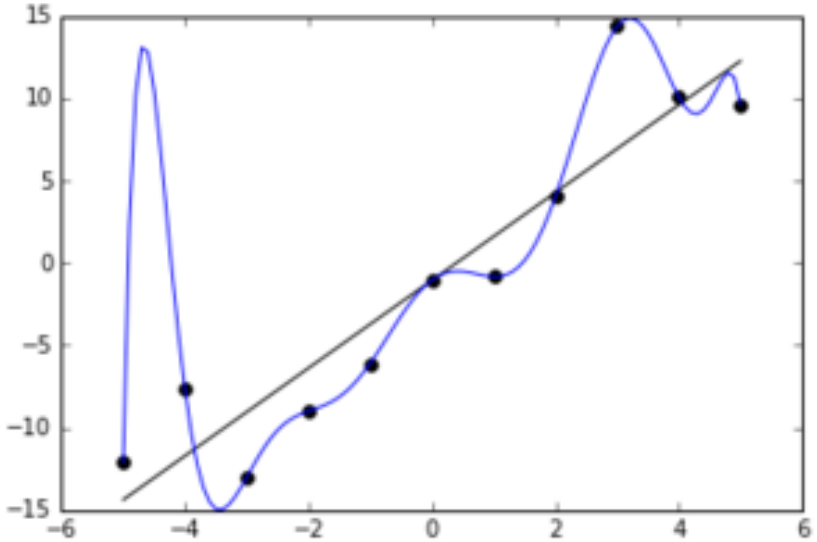
Overfitting is a key risk for machine learning

Overfitting = mistaking randomness, noise or error for real information

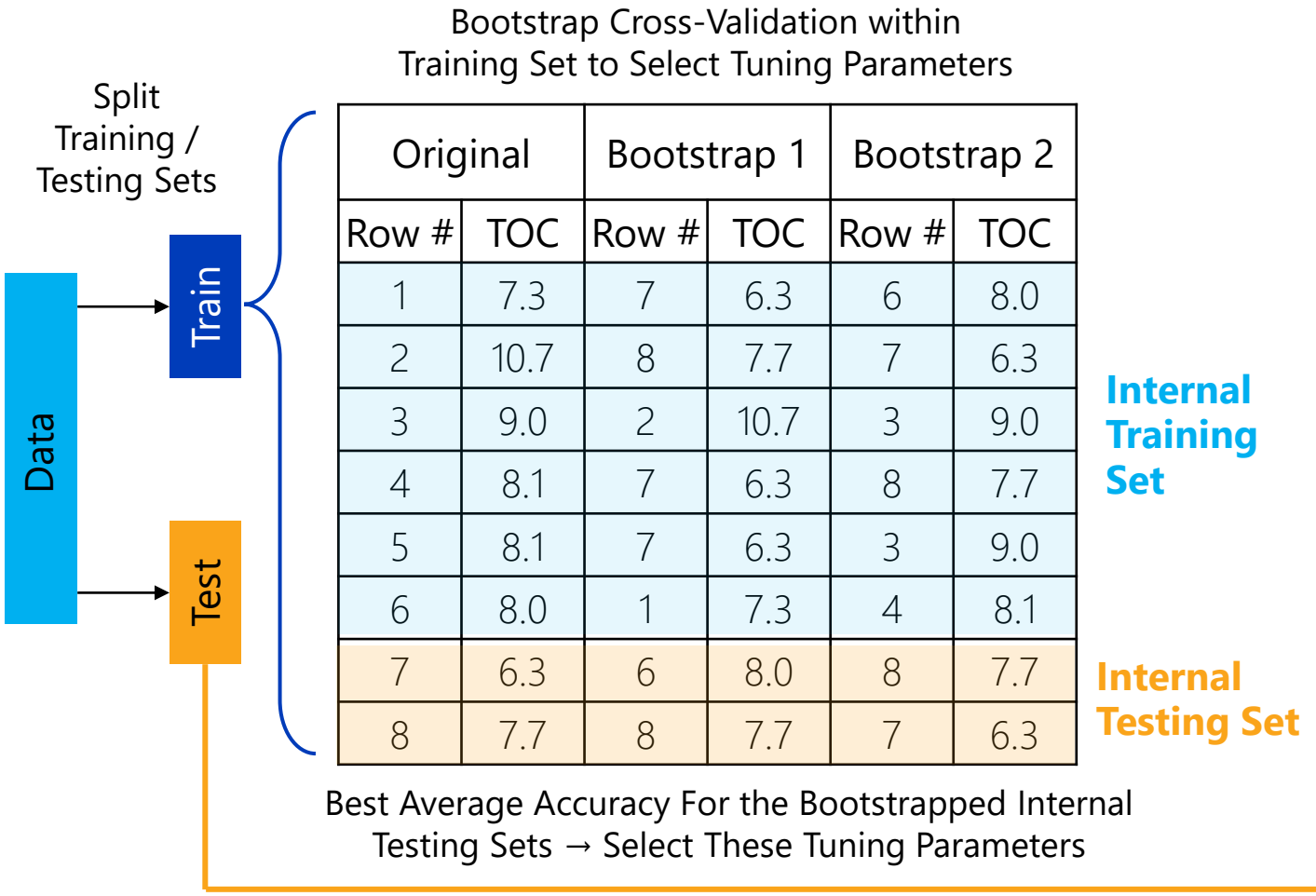
Overfit classification model



Overfit regression model



Machine learning requires training and testing to avoid overfitting



Evaluate on Testing Set

		Actual	
		Event	Normal
Prediction	Event	12	1
	Normal	16	91

4

What have been some uses of machine learning for reuse?

4A

Machine learning case study #1: Classification for fault detection

WRF Project 5048

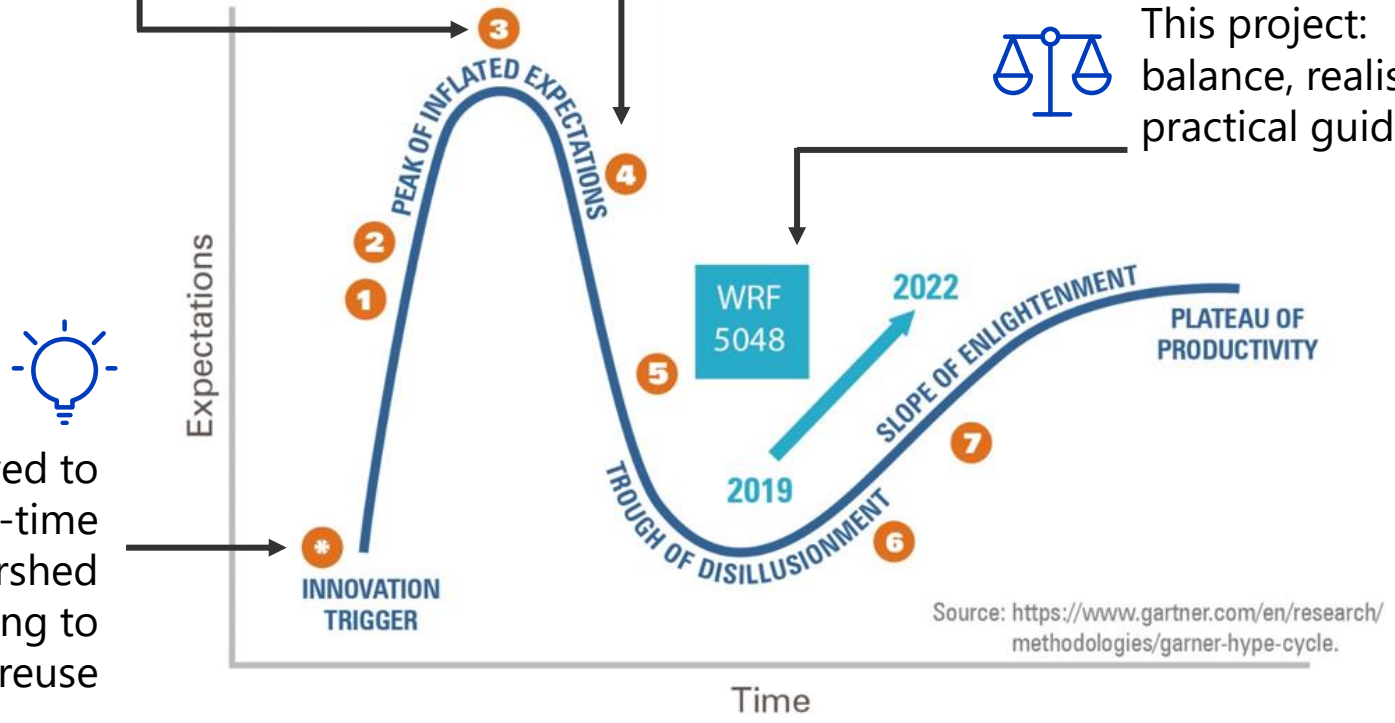
Integrating real-time collection system monitoring approaches into enhanced source control programs for potable reuse

Sensors for broader variety of water quality parameters become available (optical, specific metals, automated GC/MS)

WRF 4908 reality strikes back: maintenance, ragging, etc.

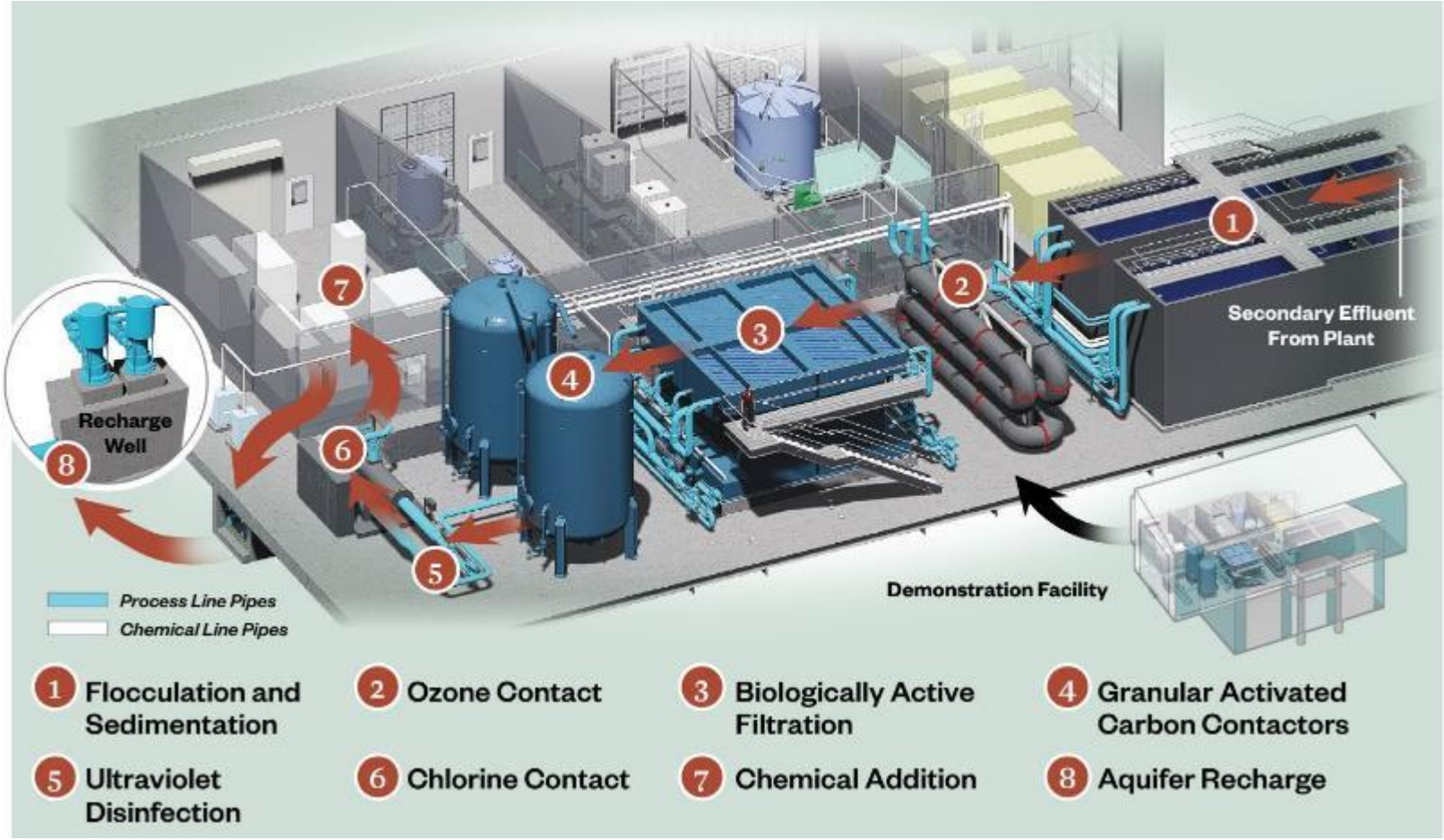
This project: balance, realism, practical guidance

Idea conceived to use real-time sewershed monitoring to protect reuse



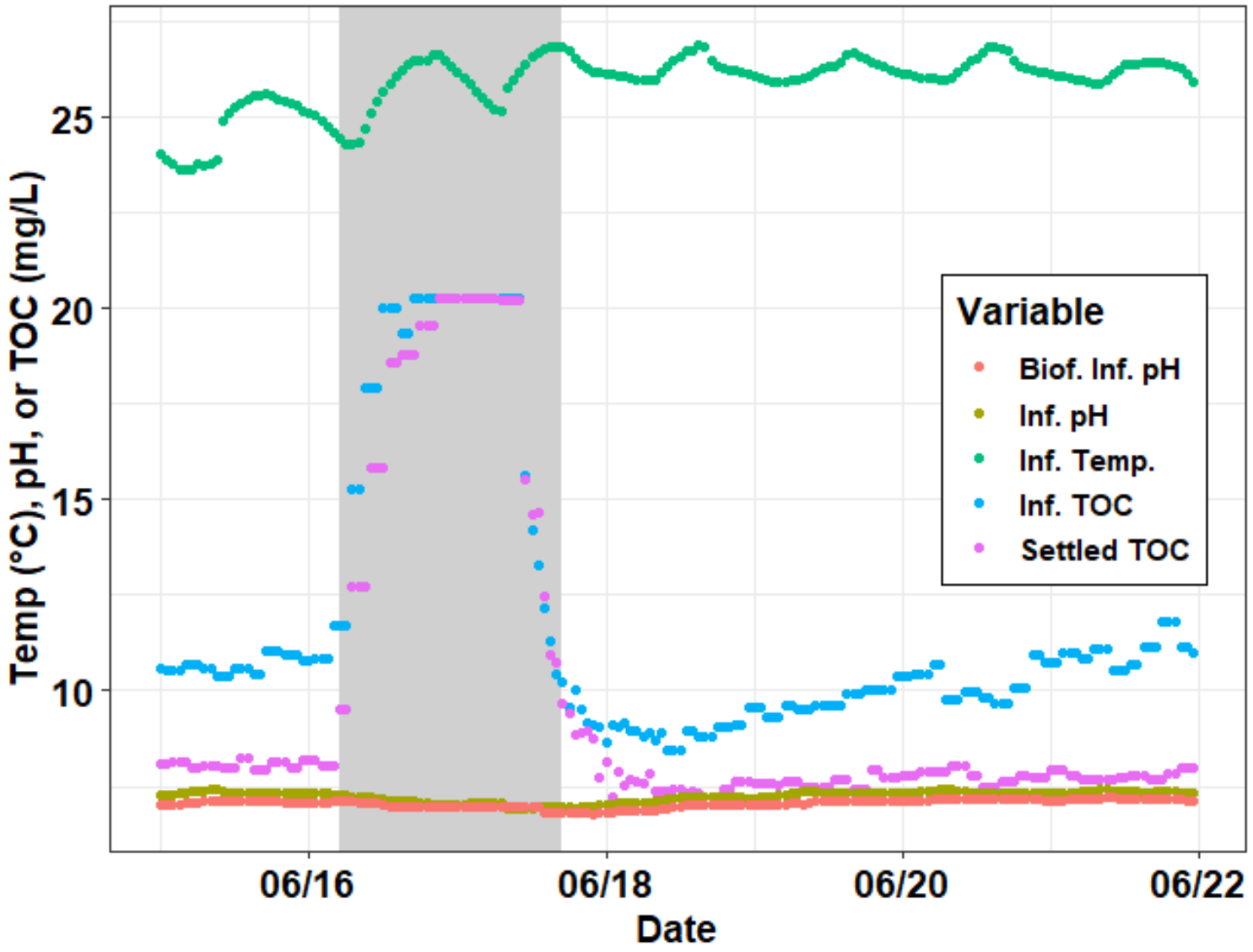
Hampton Roads Sanitation District (HRSD)'s Sustainable Water Initiative for Tomorrow (SWIFT)

- Research Center (RC) treats 1 MGD
- "Carbon-based" advanced treatment



Could machine learning be used for an alert system at SWIFT RC?

—



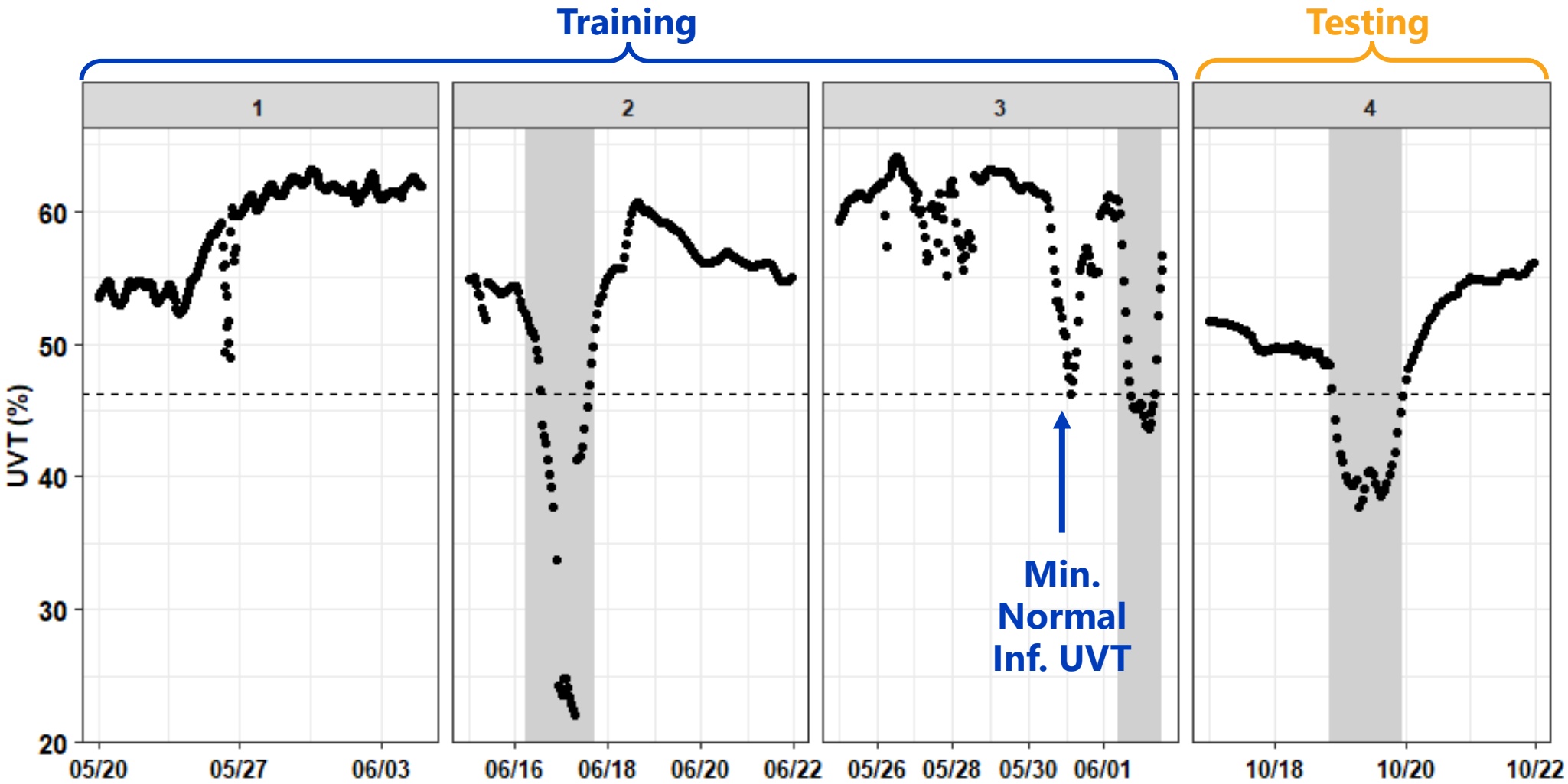
HRSD data for machine learning alert system study dataset

- **30 variables**
  - Mostly water quality but included total flow and ozone sidestream flow.
  - 1 Raw Influent
  - 13 (Plurality) Secondary Effluent
  - 8 Settled (Post Floc/Sed)
  - 4 Ozonation System
  - 4 Biofiltration Influent
- **Hourly data frequency**
- **4 time periods**
  1. May 20th, 2019 to June 4th, 2019 (Training Set)
  2. June 15th, 2019 to June 21st, 2019 (Training Set)
  3. May 25th, 2020 to June 2nd, 2020 (Training Set)
  4. October 17th, 2020 to October 21st, 2020 (Testing Set)
- **3 Industrial Discharge Events (in Time Periods #2, #3, and #4)**

Location	Variable	Units
Raw Wastewater Influent	Conductivity	**µS/cm**
Secondary Wastewater Effluent	Flow	**gpm**
	Total Nitrogen	**mg/L**
	Total Inorganic Nitrogen	**mg/L**
	Total Organic Carbon	**mg/L**
	Nitrite	**mg/L**
	Nitrogen Oxides	**mg/L**
	Nitrate	**mg/L**
	Ammonia	**mg/L**
	Conductivity	**mS/cm**
	UV Transmittance	**%**
	Turbidity	**NTU**
	pH	
	Temperature	**°C**
Settled Water	UV Transmittance	**%**
	Monochloramine	**mg/L**
	Ammonium	**mg/L**
	Total chlorine	**mg/L**
	Redox potential	
	Total Organic Carbon	**mg/L**
	Total Nitrogen	**mg/L**
	Free Ammonia	**mg/L**
Ozonation System	Ozone Dose	**lbs/day**
	Ozone Sidestream Flow	**gpm**
	Ozone Residual Setpoint	
	Ozone Residual	**mg/L**
Biofiltration Influent	UV Transmittance	**%**
	Total Chlorine	**mg/L**
	pH	
	Redox potential	mV

Benchmarking against fixed thresholds

Threshold set to maximum or minimum normal value within training set



Screening results

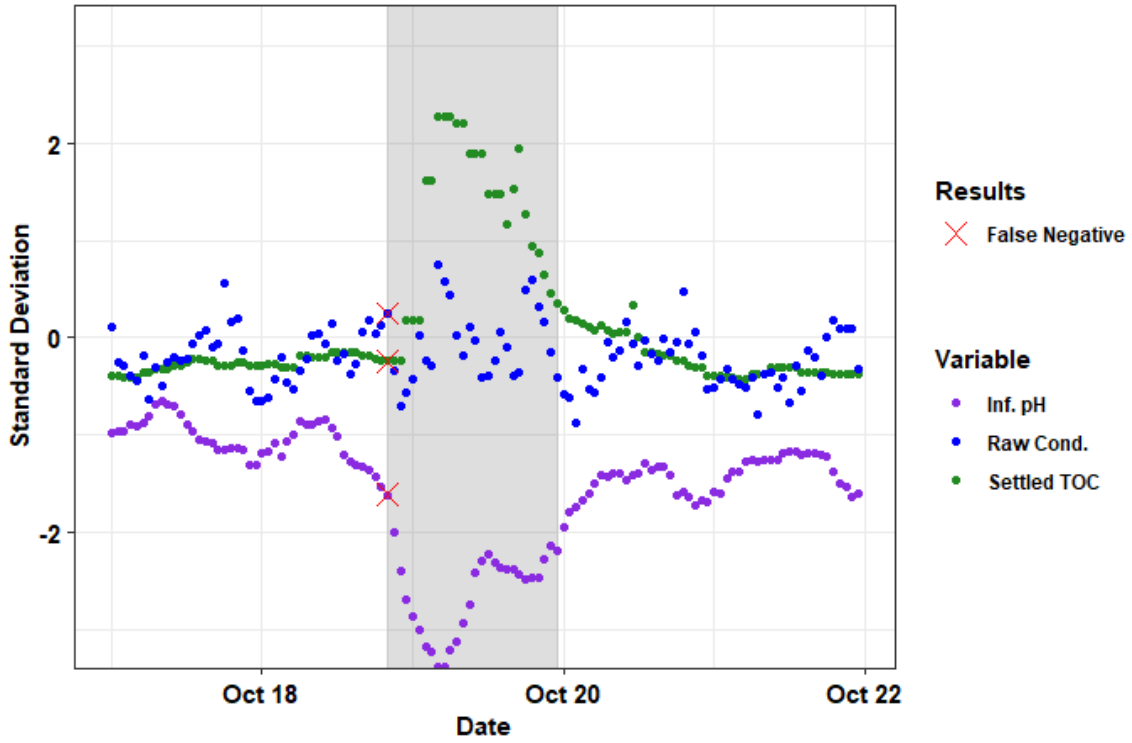
- **35 models screened**
- **Six selected for in-depth evaluation**
  - » Two highest training set accuracy
  - » Two highest test set accuracy
  - » Two highest event sensitivity
    - Event sensitivity = predicted event when true answer was event

Model	Abb.	Training Set Accuracy	Testing Set Accuracy	Event Sensitivity
Boosted Tree	bstTree	99%	95%	**89%**
Cost-Sensitive C5.0	C5.0Cost	99%	**97%**	86%
Oblique Random Forest with Support Vector Machines	ORFsvm	99%	**96%**	82%
Penalized Logistic Regression	plr	**100%**	88%	50%
Random Forest Rule-Based Model	rfRules	98%	54%	**100%**
Support Vector Machines with Radial Basis Function Kernel	svmRadial	**99%**	83%	25%

1st highest out of 35 2nd highest out of 35

Boosted Tree (bstTree)

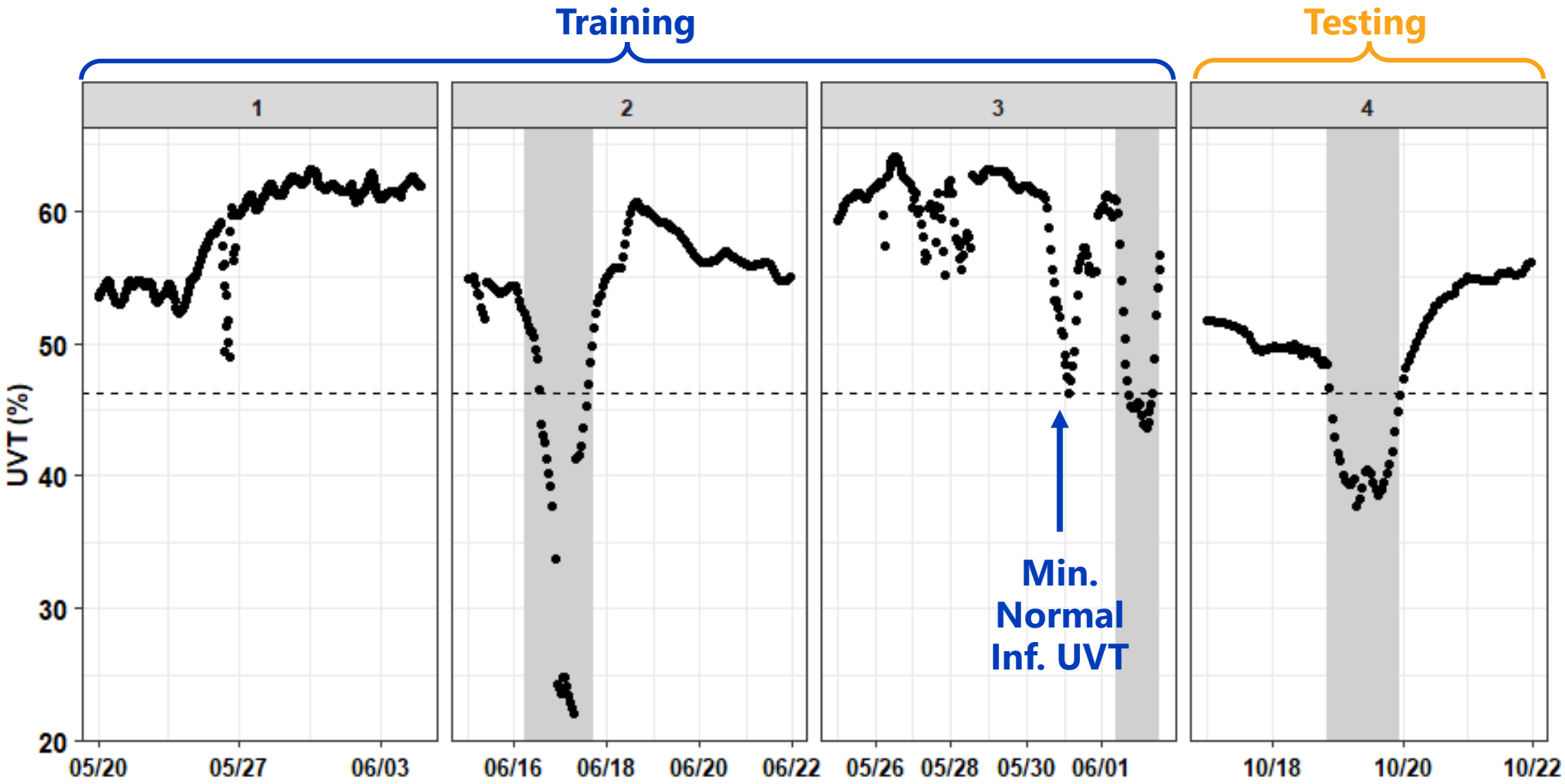
- Boosted Tree (bstTree) was the most accurate model after optimization
- Testing set accuracy = 99.3%
- 1 false negative, 0 false positives
- Would have detected event in about 1 hour



Training Set Accuracy	99.3%
Testing Set Accuracy	99.2%
Testing Set Balanced Accuracy	98.2%
Testing Set Cohen's Kappa	0.976
Testing Set Event Sensitivity	96%
Testing Set False Positives	0
Time until 1st Detection (hr)	1
Preprocessing	None
Variables Used	13
Tuning Parameters	maxdepth=3, nu=0.1, mstop=150

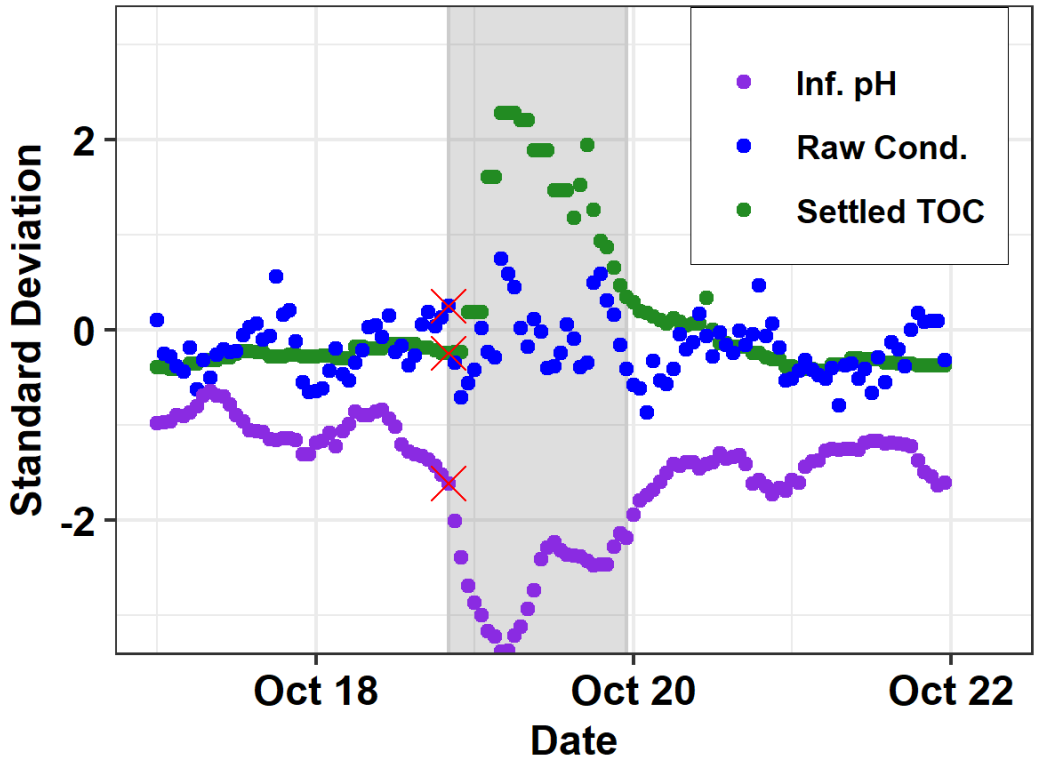
Fixed thresholds results

- Among Fixed Thresholds, Influent UVT had highest test set accuracy, 98.3%
- Only one more error compared to best Machine Learning model



Conclusions

- The best machine learning model (bstTree) performed better than any fixed threshold with a **testing set accuracy of 99.3%**.
- **Zero false positives** over 5-day testing set.
- Time until first detection ~1 hour.
- However, only better than Inf. UVT threshold by about 1% with the data used for this proof-of-concept.
- More data (especially if included 4+ industrial events) would improve both:
  - The accuracy of the machine learning models
  - The confidence of the comparative evaluation



For more information:

Thompson, K.A., Branch, A., Nading, T., Dziura, T., Salazar-Benites, G., Wilson, C., Bott, C., Salveson, A., Dickenson, E., 2022. Detecting Industrial Discharges at an Advanced Water Reuse Facility Using Online Instrumentation and Supervised Machine Learning Binary Classification. Front. Water 4, 1014556.

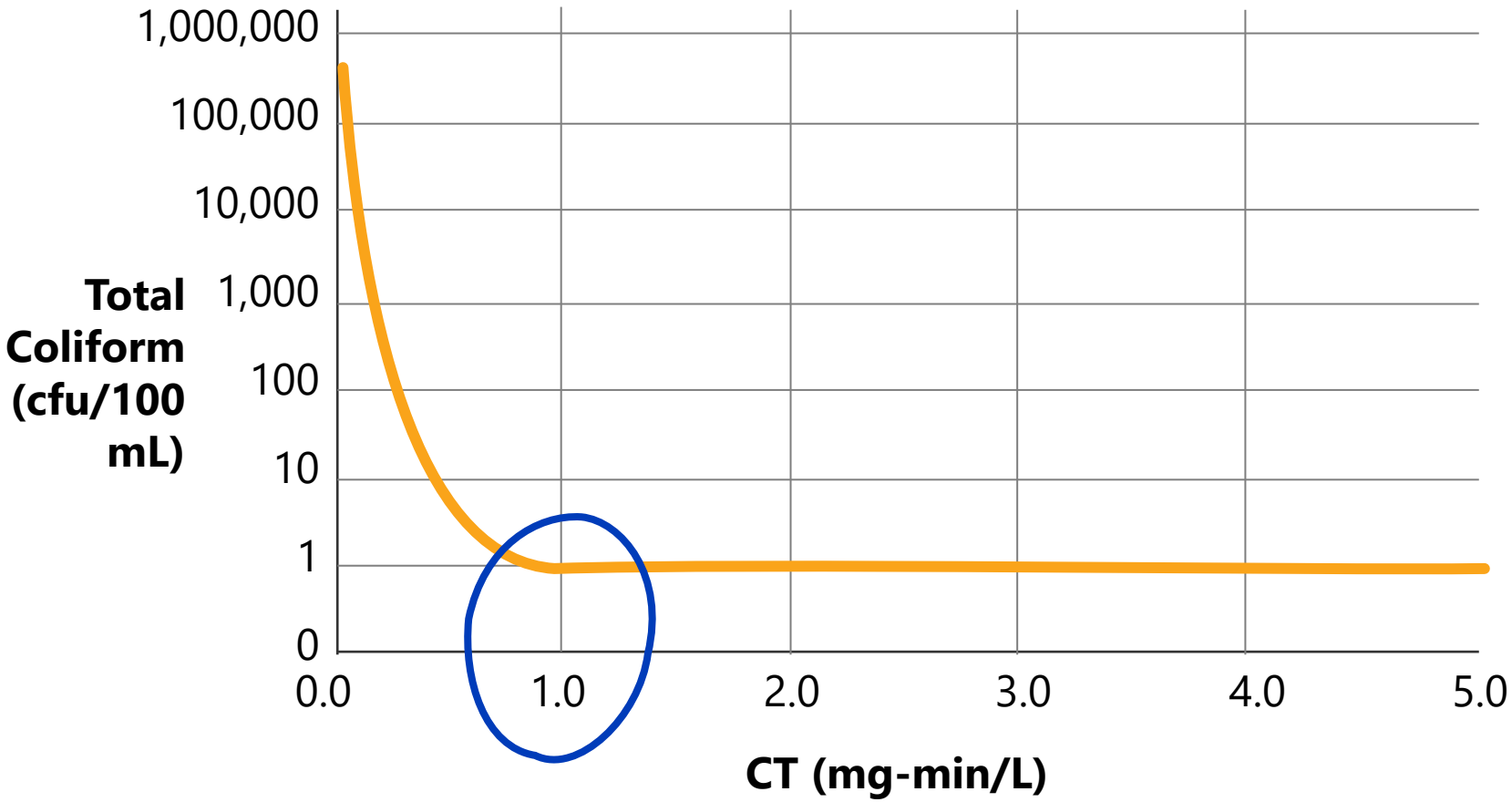
4B

Machine learning case study #2: Regression for soft sensor

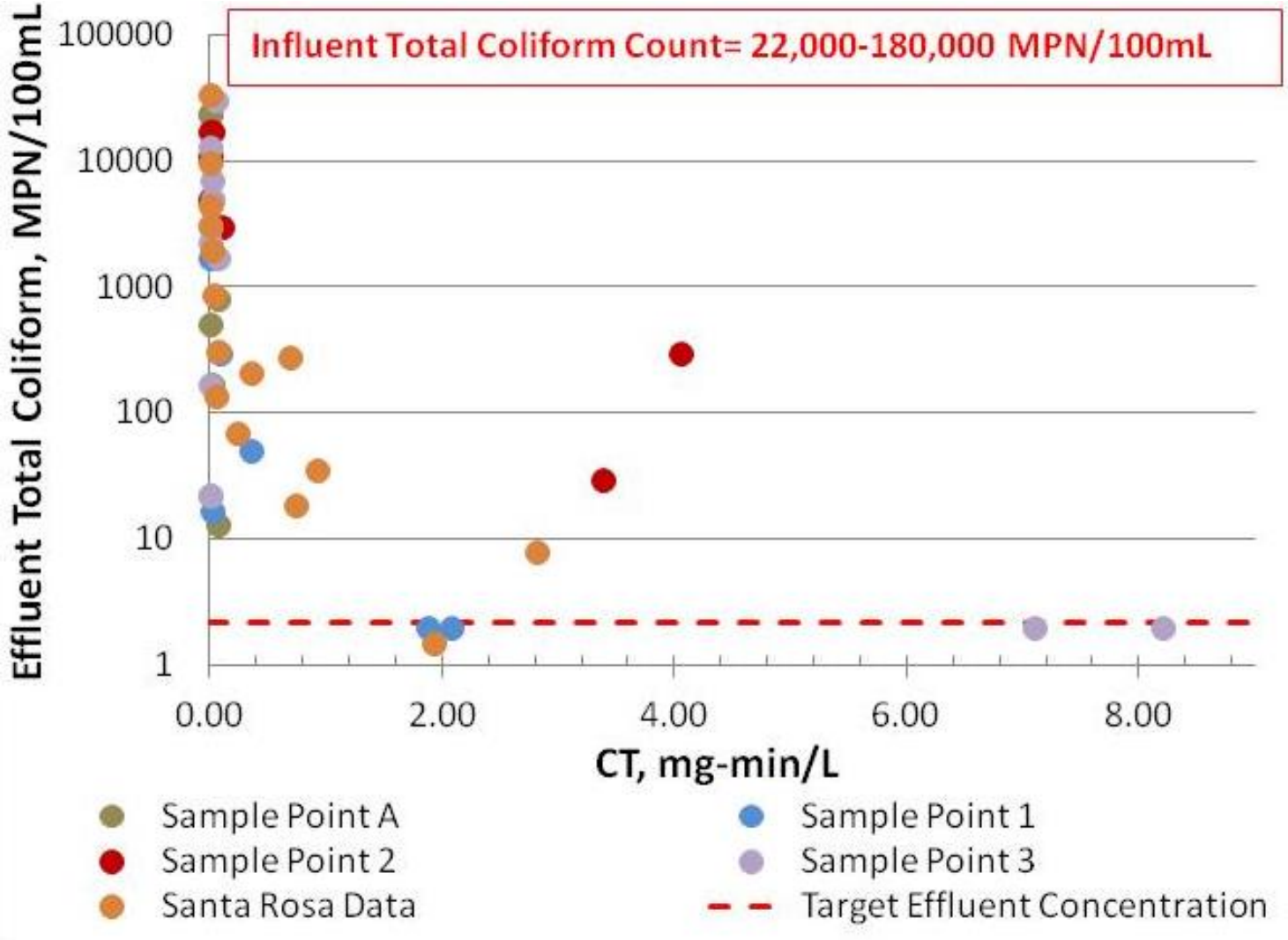
WRF 5129

Demonstration of innovation to improve pathogen removal, validation, and/or monitoring in Carbon Based Advanced Treatment (CBAT) for potable reuse

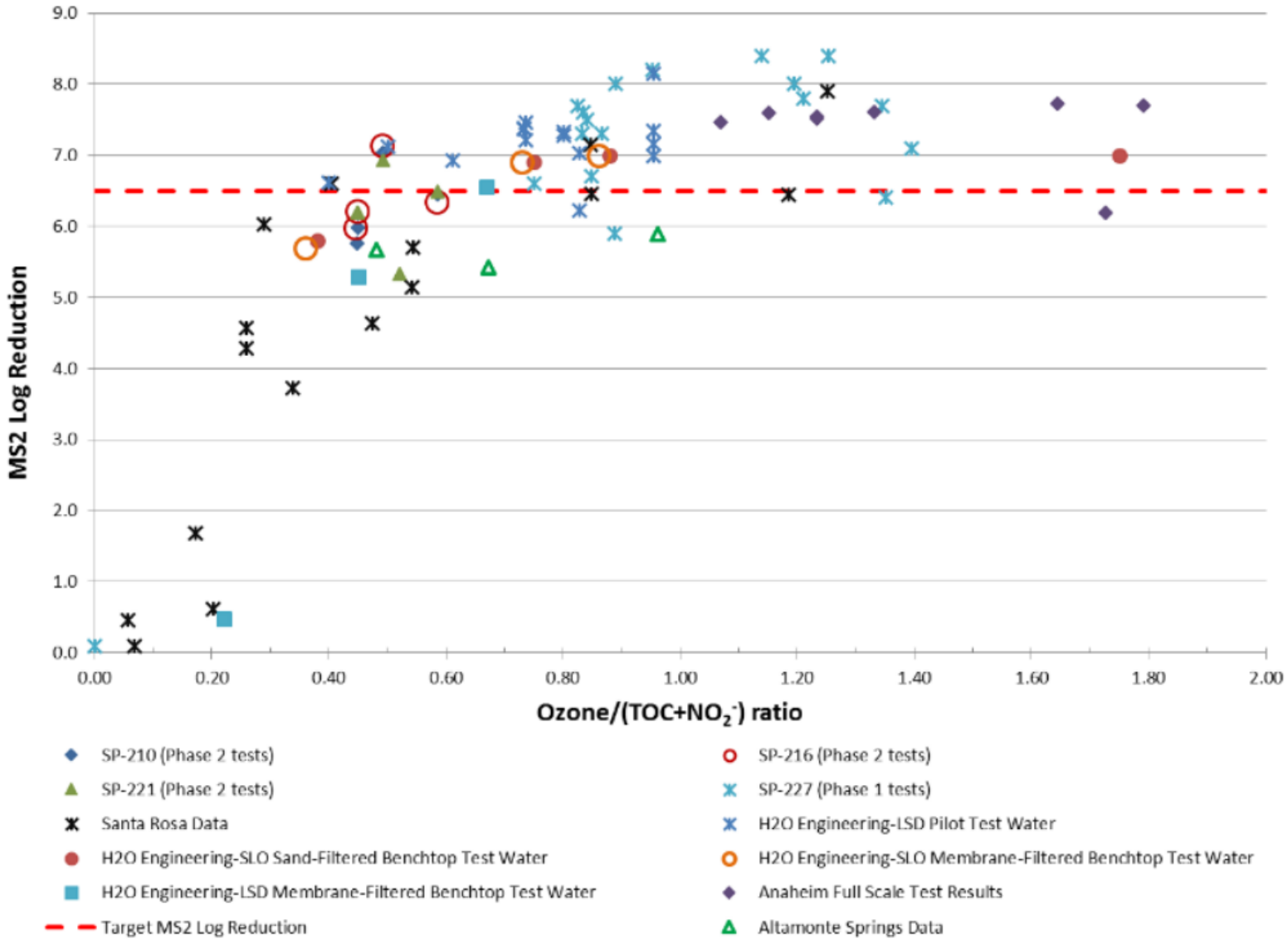
Current regulatory paradigm: Disinfection = Concentration × Time (CT)



Not an ideal relationship



O_3:(TOC+NO_2) – a better way to dose ozone



One disadvantage – TOC measures more slowly than ozone residual

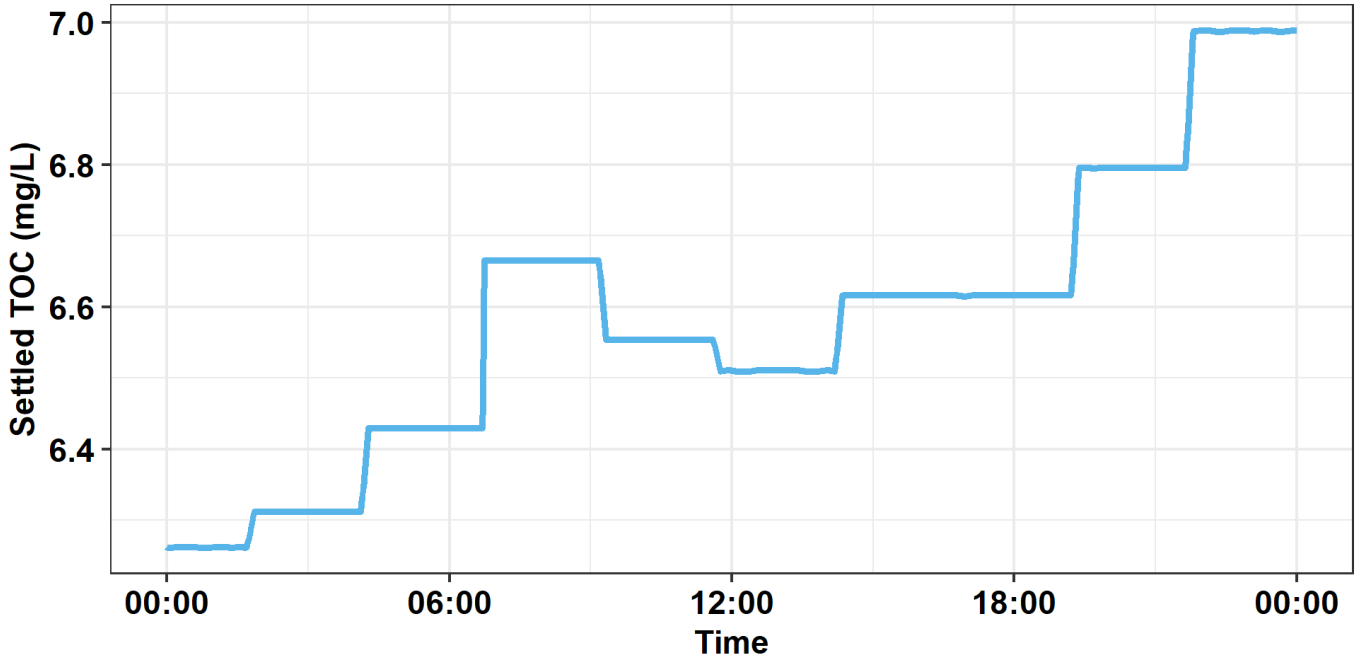
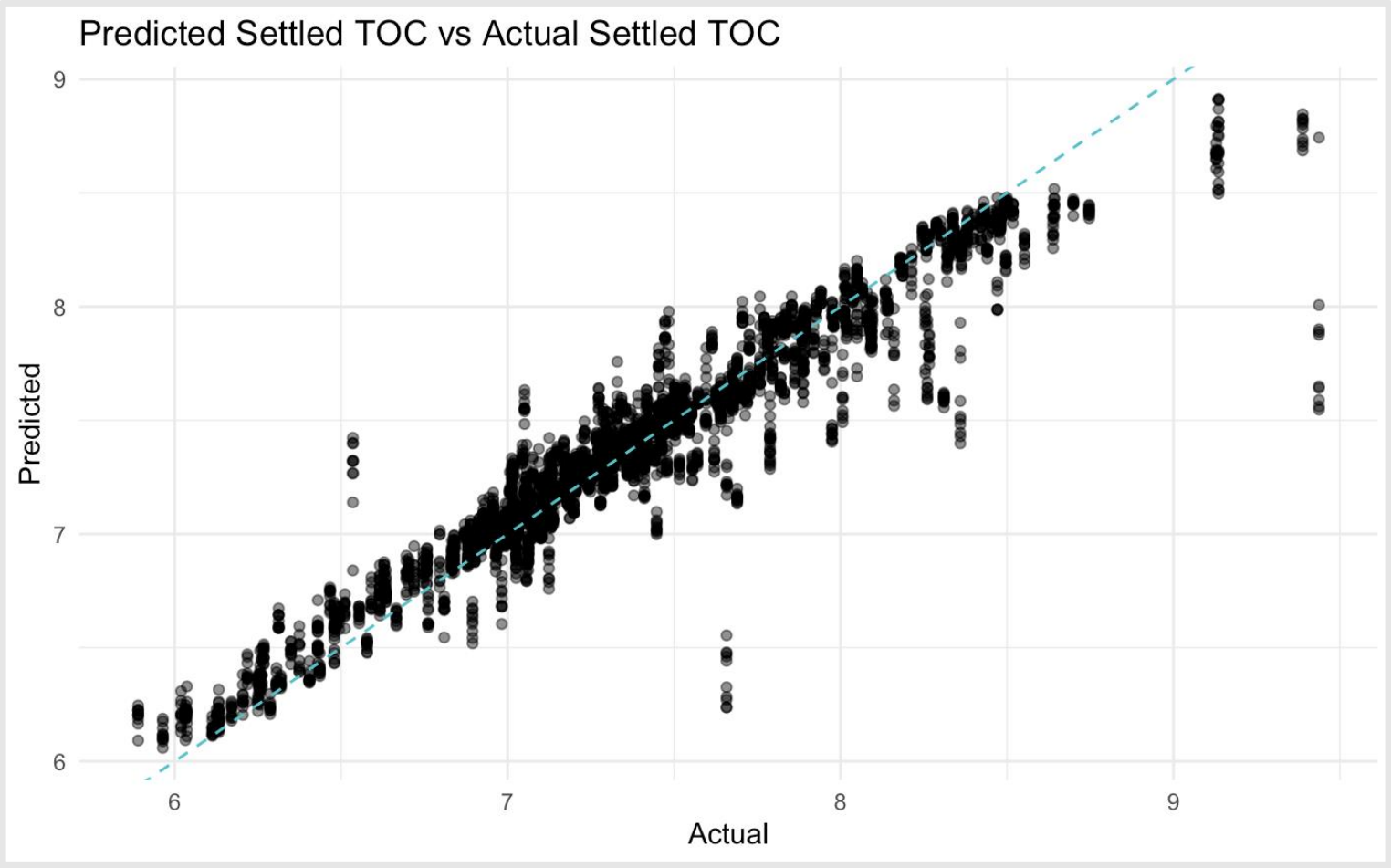


Table 1. General characteristics of tested ozone analyzers.

Parameter	Meter A: Kuntze Krypton Dis	Meter B: Rosemount 499A OZ	Meter C: Hach CL17	Meter D: 2B Technologies UV-106-W
Measurement range	0–5 mg/L	0–3 mg/L	0–5 mg/L	0–100 mg/L
Analysis time	Continuous	Continuous	2.5 min	10 s

Chen, E.C., et al. 2020. Ozone Sci. Eng. 42, 213–229.

ML predicts TOC to enable rapid TOC-based ozone dosing



- Random Forest
- Root Mean Square Error = 0.171 mg/L

5

How should we go about applying ML/AI?

Have a clear problem statement

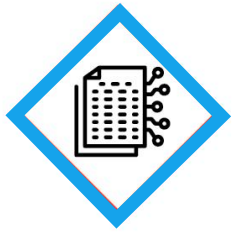
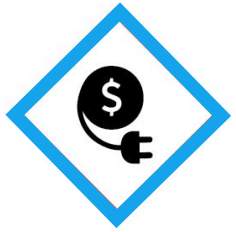
- What are you trying to predict?
- How will that prediction be used?
- What information would feed the prediction?
  - » Do you have that information?
- Would a simpler tool get the job done?

Yishi Zuo, 2021

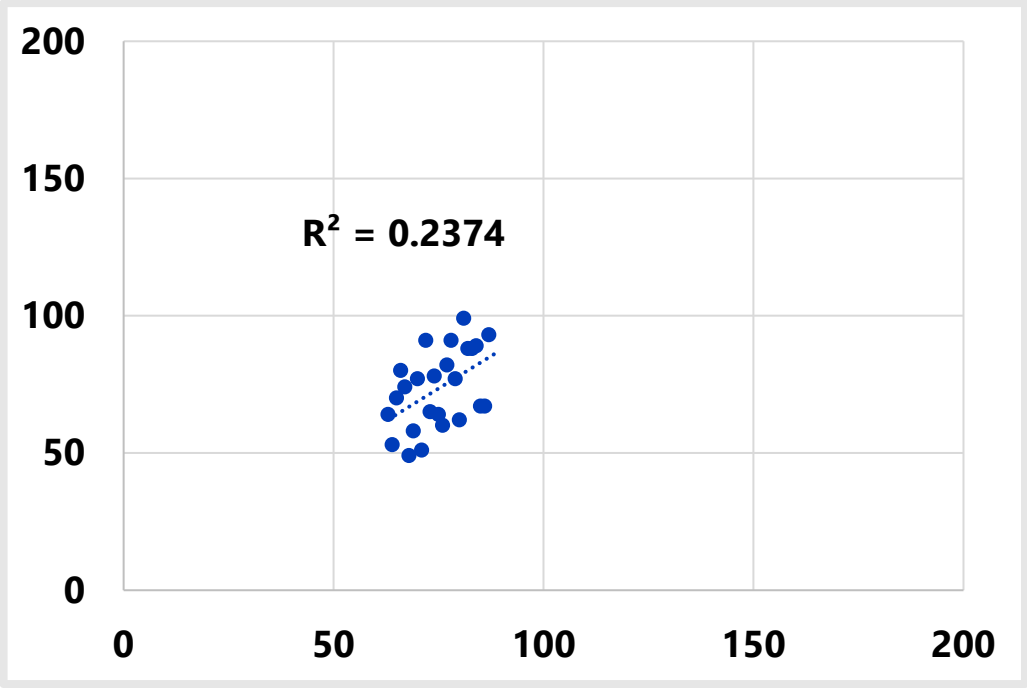
Establish appropriate benchmarks, metrics, and goals

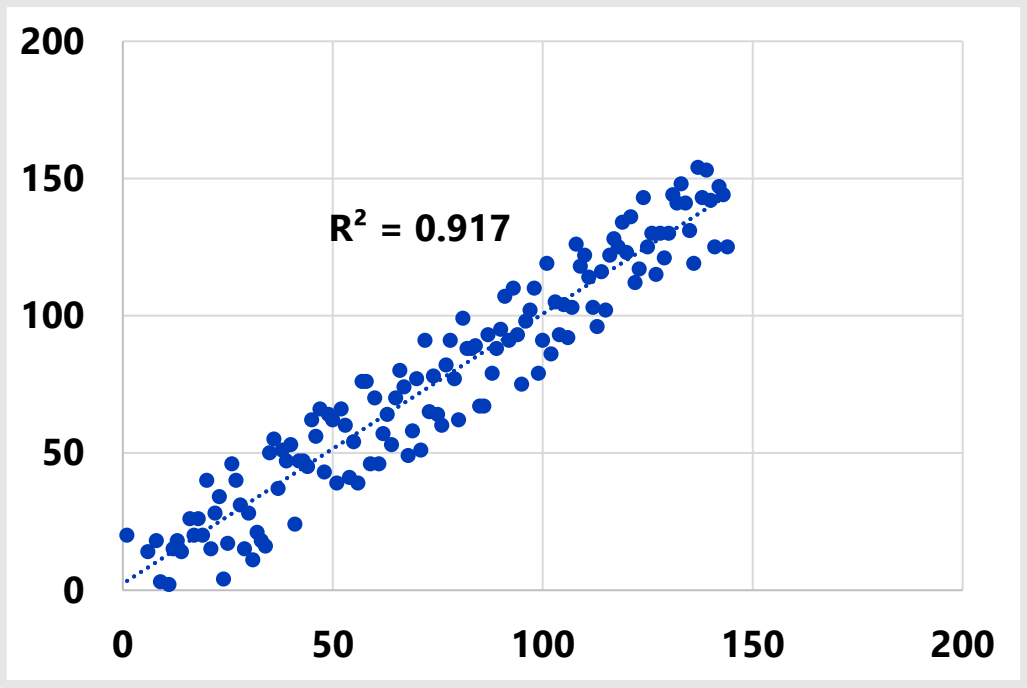
XKCD.COM PRESENTS A NEW "IS IT CHRISTMAS" SERVICE TO COMPETE WITH ISITCHRISTMAS.COM

ML is best applied to data rich environments and complex problems

	**Data Variability** Does typical pre-existing data (both influent water quality and operational settings) provide enough range for meaningful model training?		**Data Availability** Does typical pre-existing instrumentation provide enough input variables for a ML model? Has the data been collected for long enough or at a high enough frequency to generate a sample size that is valid for ML?		**Process Complexity** Is the process sufficiently complex, non-linear, or unknown that it is reasonable to expect that an ML model would meaningfully outperform a simpler (e.g. linear) model?
	**Process Flexibility** Do operators have the power to adjust operation settings, or is operation constrained by design or regulation?		**Cost & Energy** Does the process represent a substantial overall fraction of the energy usage or operational costs at utilities? Would efficiency gains in this process meaningfully impact the overall energy or expenses for the facility?		**Process Dynamicity** Do the input variables or ideal response vary quickly enough to merit an automated or real-time ML implementation?

ML modeling requires not only high sample size, but also meaningful range

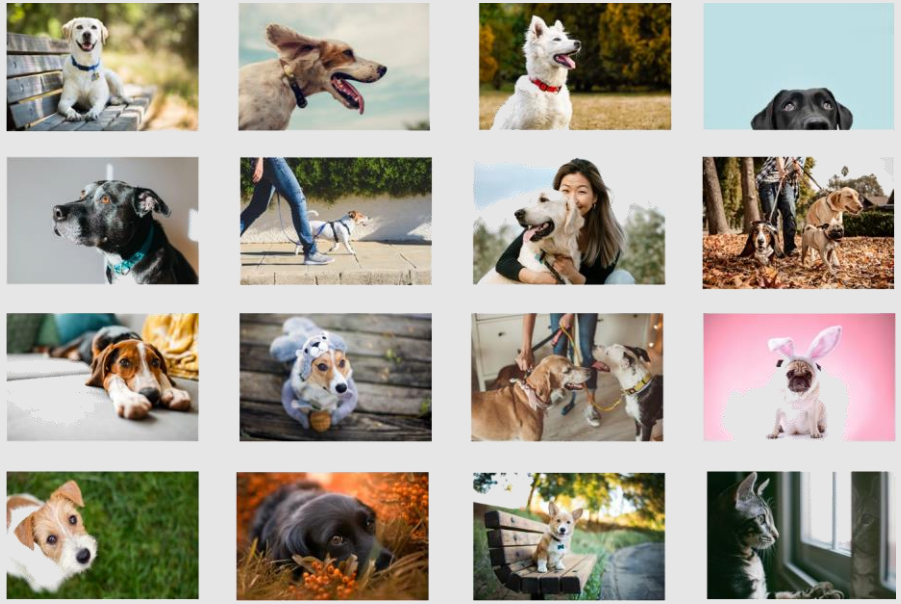




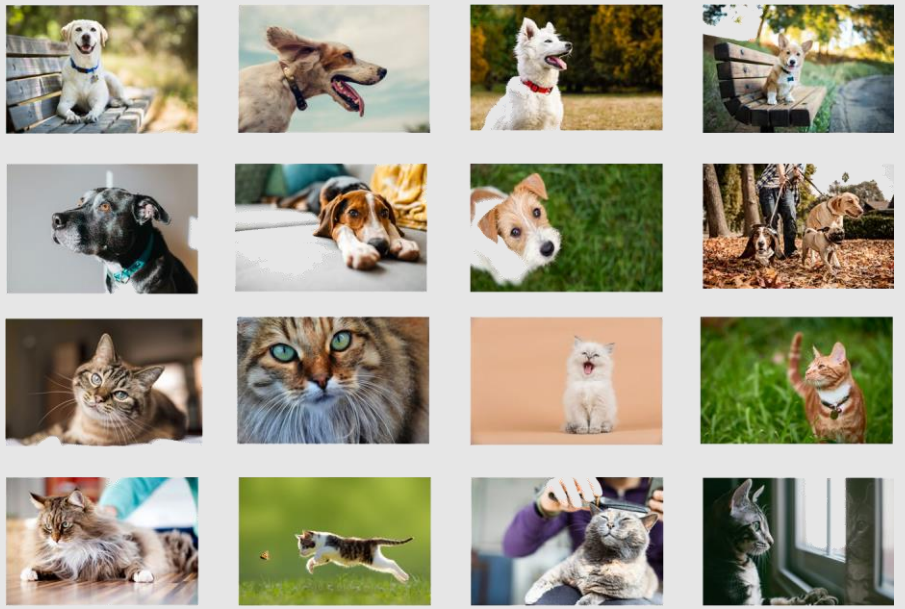
ML modeling requires not only high sample size, but also meaningful range

Classification: Dog or Cat?

Unbalanced Training Set

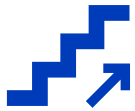


Balanced Training Set

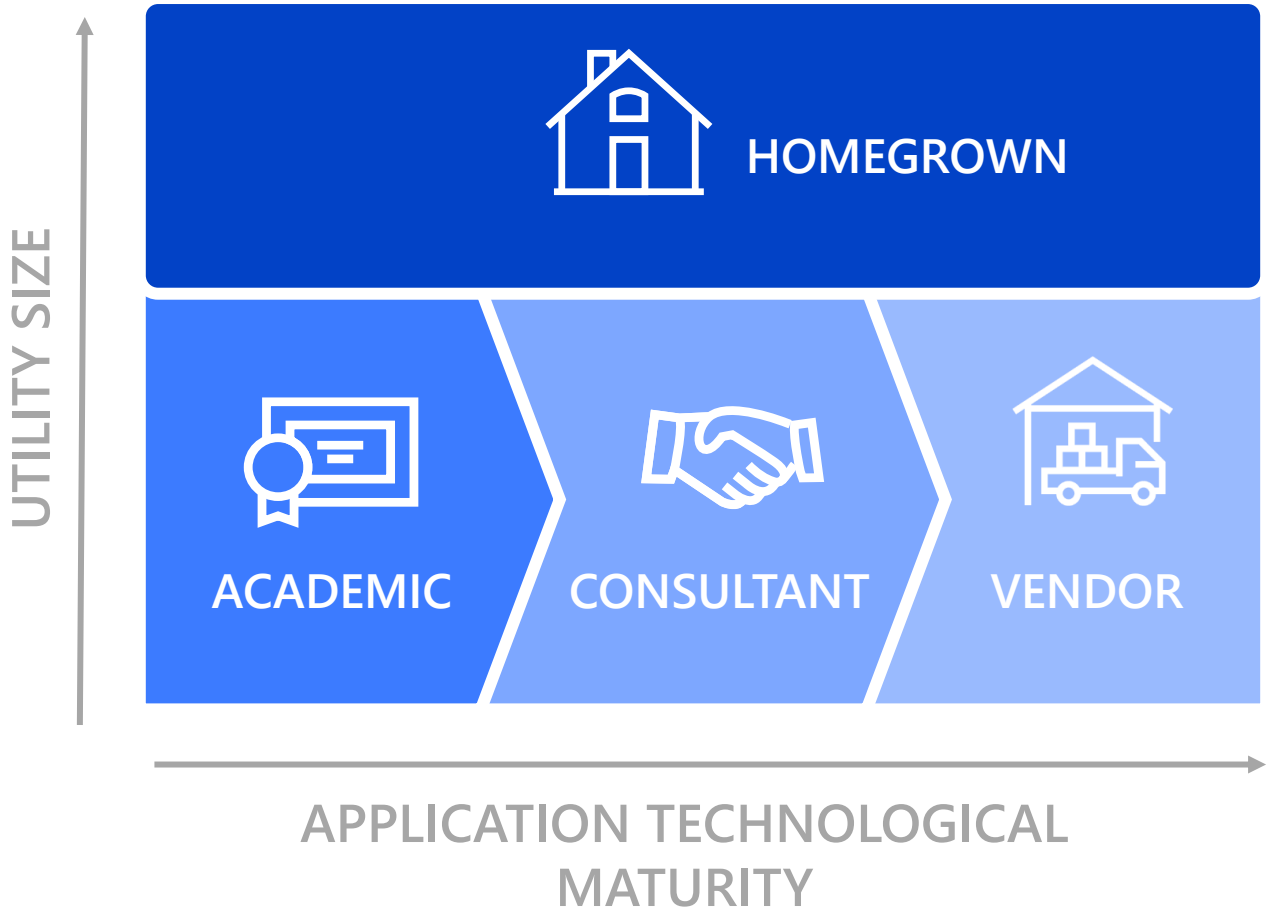


There are several partnership strategies through which to implement ML/AI

Factors such as transparency, cost, and scalability impact ML/AI implementation decisions

Transparency	Do I want to see the code?	
Cost	Subscription fees? In-house employees(s)?	
Scalability	How quickly can this go from trial to full-scale?	
Long-Term Sustainability	Stable costs? Employee turnover?	
Cybersecurity	Are sufficient firewalls in place?	

The best way to begin and implement ML/AI depends on factors such as utility size



Acknowledgements

- **Carollo Team:** Andy Salveson, Amos Branch, Claire Teng (Baylor University Intern)
- **Collaborating Organizations:** Hampton Roads Sanitation District, Southern Nevada Water Authority, Baylor University, Jacobs
- **Funding:** Water Research Foundation (Projects #5048 and #5129)

Thank You!

kthompson@carollo.com