the international water association

The Application of Artificial Intelligence (AI) in Water and Wastewater Treatment

16 APRIL 2026

inspiring change

WEBINAR INFORMATION – HOUSEKEEPING RULES

- This webinar will be **recorded and made available "On-demand"** on the Connectplus platform, with relevant information.
- The **speakers** are responsible for **securing copyright permissions** for any work that they will present of which they are not the legal copyright holder.
- The opinions, hypothesis, conclusions or recommendations contained in the presentations and other materials are the **sole responsibility of the speaker(s)** and do not necessarily reflect IWA opinion.

WEBINAR INFORMATION – HOUSEKEEPING RULES

- **'Chat' box**: please use this for general requests and for interactive activities.
- **'Q&A' box**: please use this to send questions to the panelists. (We will answer these during the discussions)

Please Note: *Attendees' microphones are muted. We cannot respond to 'Raise Hand'.*

AGENDA

- **Introduction and Housekeeping**
  Wen Ma, Université de Sherbrooke, Canada ; Yumeng Zhao, Harbin Institute of Technology, China

- **Presentation 1: Machine learning supporting the operation of wastewater treatment processes**
  Peter Vanrolleghem, Universite Laval, Canada

- **Presentation 2: Auto-AMP: AI for water Infrastructure Management**
  Nicholas St-Gelais, CANN Forecast, Canada

- **Presentation 3:Understanding Micropollutant Rejection Mechanism by Polyamide Membranes via Data-Knowledge Co-Driven Machine Learning**
  Ruobin Dai, Tongji University, China

- **Q & A including Panel Discussion**
  All speakers and moderators

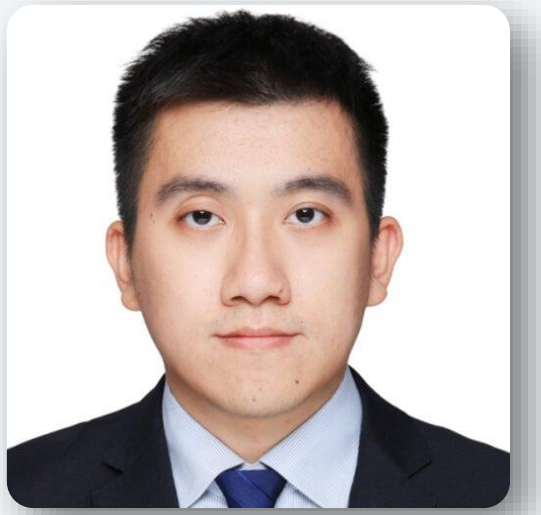
- **Closing remarks**
  Wen Ma, Université de Sherbrooke, Canada ; Yumeng Zhao, Harbin Institute of Technology, China

The Application of Artificial Intelligence (AI) in Water and Wastewater Treatment

IWA WEBINAR | 16 April 2026 | 9:00 ET

Peter Vanrolleghem
Université Laval
Canada

Nicolas St-Gelais
CANN Forecast
Canada



Ruobin Dai
Tongji University
China

Register

https://www.iwa-network.org/learn/the-application-of-artificial-intelligence-ai-in-water-and-wastewater-treatment-1

Moderator

Wen Ma, Université de Sherbrooke, Canada
Yumeng Zhao, Harbin Institute of Technology, China

Co-organizer

IWA the international water association — DESIGN, OPERATION AND MAINTENANCE OF DRINKING WATER TREATMENT PLANTS

PRESENTATION 1

Machine learning supporting the operation of wastewater treatment processes: Get ready – the future is near!

the international water association

Peter Vanrolleghem

Full Professor
Université Laval, Canada

Peter Vanrolleghem obtained his degrees in Bio-engineering and PhD in Environmental Technologies from Ghent University (Belgium).

In 2006 he immigrated in Quebec, as a Canada Research Chair in Water Quality Modelling. His multi-cultural research team, **model*EAU***, focuses on urban wastewater systems, including nutrient removal/recovery, micropollutants, pathogens, and greenhouse gas emissions, using modelling, monitoring, and process control.

He has served on the Boards of the Water Environment Federation and the International Water Association, and currently chairs IWA Publishing. He founded CentrEau in Québec and received the 2023 NSERC Donna Strickland Prize for high-impact research.

the international water association

Machine learning supporting the operation of wastewater treatment processes:

Get ready – the future is near!

PETER A. VANROLLEGHEM

inspiring change

UNIVERSITÉ LAVAL

Machine learning supporting the operation of wastewater treatment processes:

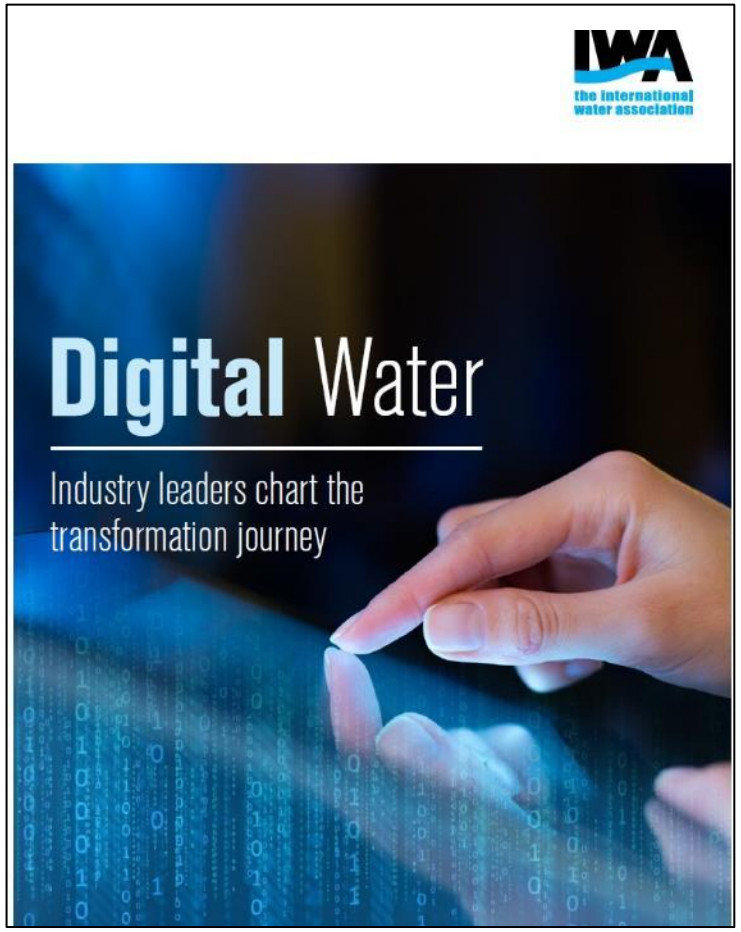
Get ready – the future is near!

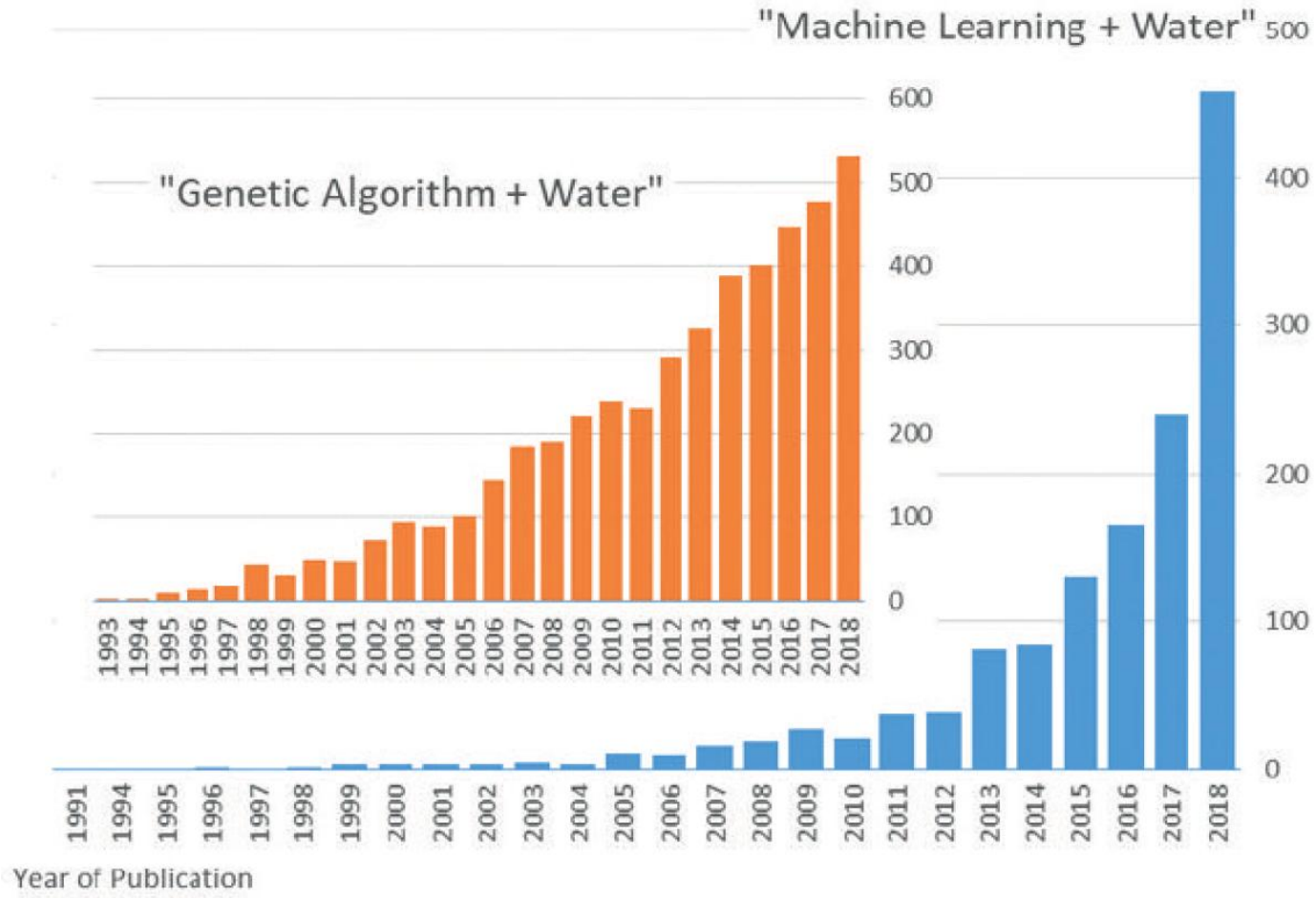
PETER A. VANROLLEGHEM

NATHALIA COVRE, MOSTAFA KHALIL, FEIYI LI, BERNARD PATRY, MARCELLO SERRAO, JEFF SPARKS, JEAN-DAVID THERRIEN

inspiring change

Digital transformation of the water sector





UNIVERSITÉ LAVAL

Rate of Digital Transformation in the Water Sector

- Changes to water systems are (frustratingly) slow
  - Installed (legacy) base has life expectancy of 20, 50, 100 years
  - Typical time constants for water system change are 10-20 years…
  - Digital transformation has time constant of 1-3 years

- For certain aspects of digitalization, this is maybe a good characteristic
  - Security issues
  - Ethical issues

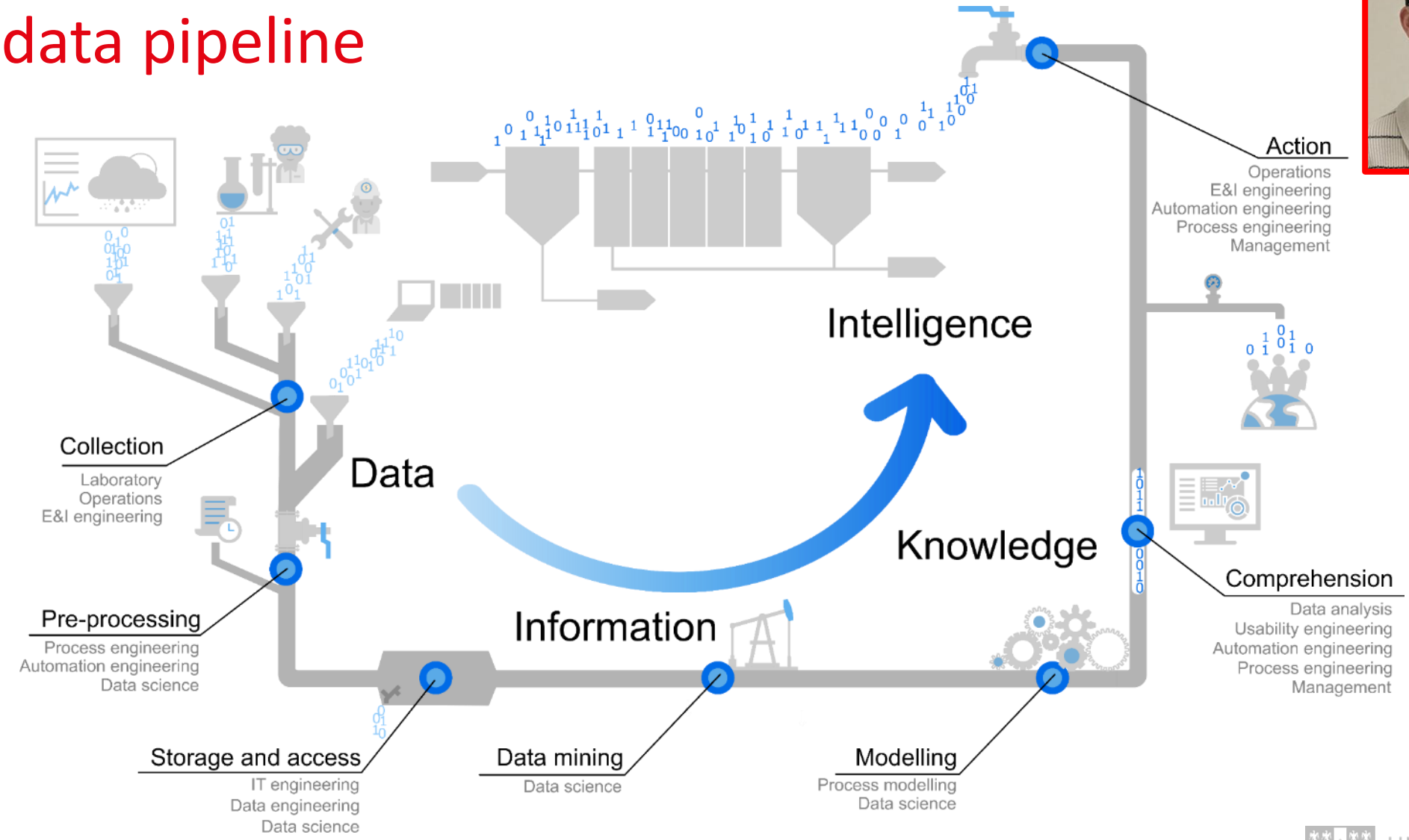
Rate of Digital Transformation in the Water Sector

- However, one important paradigm shift is occurring:
  - Before the digitalization hype:
    - Utility management was not really into ICA, modelling, …
    - Many good ideas never made it to practice **(bottom-up)**
  - With the digitalization hype:
    - Buy-in from utility management
      - CEO's and utility boards (Mayors!) now lead the way **(top-down)**
      - Chief Digital Officers (CDO) are installed to make it happen, **quickly**
      - Personnel is empowered to develop new approaches

"If you have any doubt, just try it"

UNIVERSITÉ LAVAL

The data pipeline



A critical review of the data pipeline: how wastewater system operation flows from data to intelligence.
Therrien J.D., Nicolai N., Vanrolleghem P. A. (2020) Water Science & Technology. DOI: 10.2166/wst.2020.393

UNIVERSITÉ LAVAL

Data quality control

- Wastewater is nasty

- Noise
- Outliers

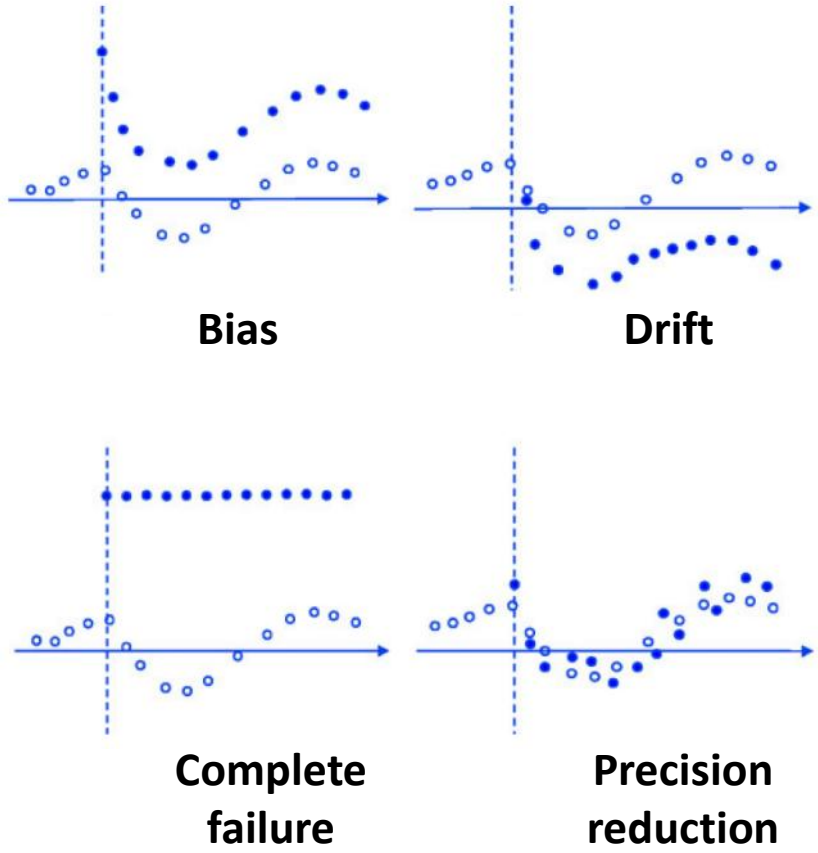
- Faults
- Gaps in data

Data quality control

- Wastewater is nasty

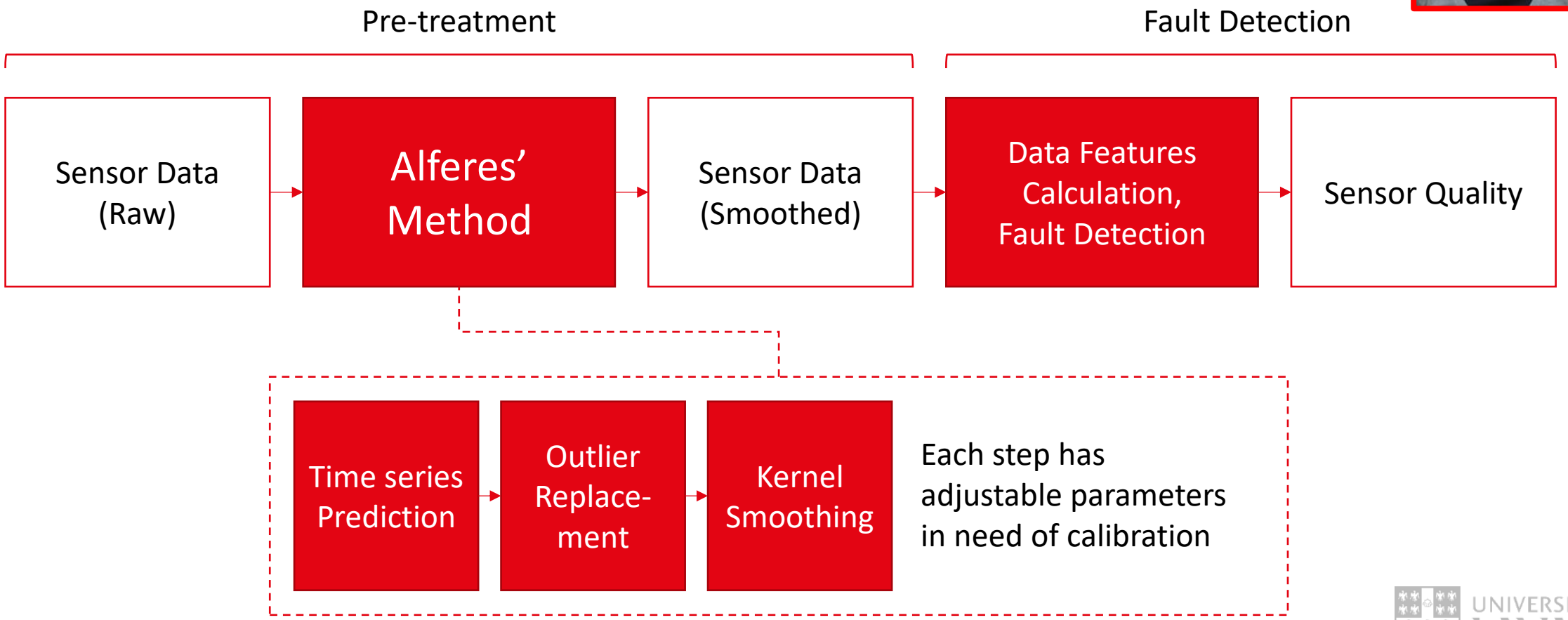
Data quality control – Sensors are not perfect

- Sensors fail or need maintenance during operation
- Collected data need validation in real time
- High frequency sensors are hard to deal with, due to the high volume of data generated



Data quality control – Sensors are not perfect

- Previous work – Alferes' method

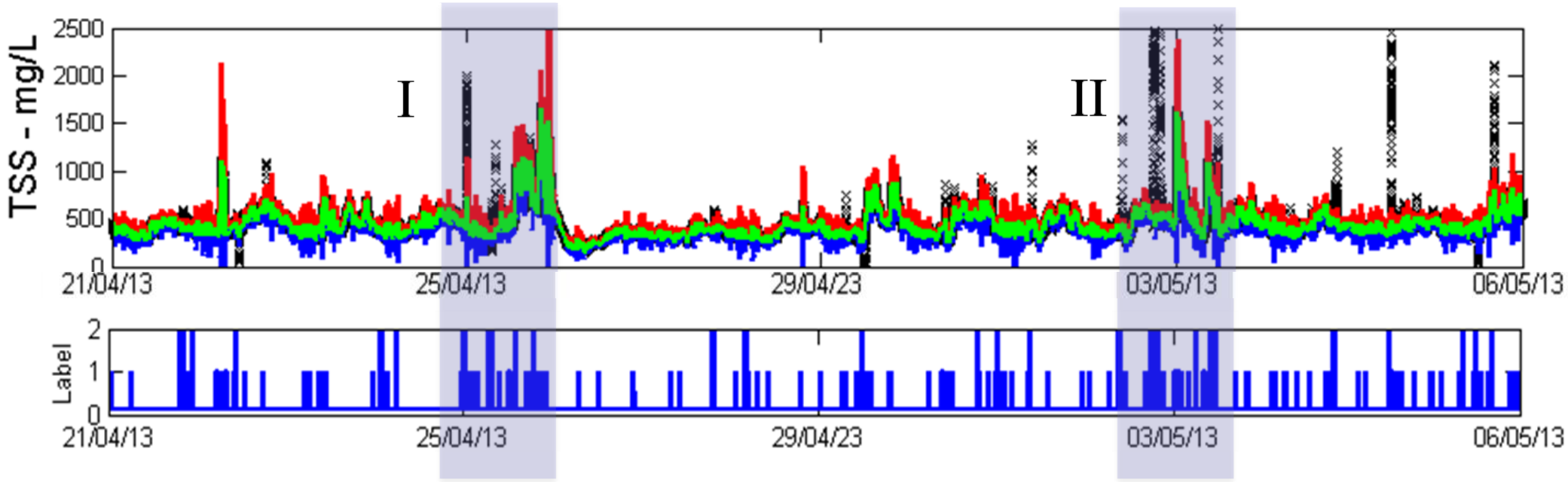


Alferes J. and Vanrolleghem P.A. (2016) Efficient automated quality assessment: Dealing with faulty on-line water quality sensors. AI Communications, 29, 701-709.

UNIVERSITÉ LAVAL

Data quality control – Sensors are not perfect

- Previous work – Alferes' method

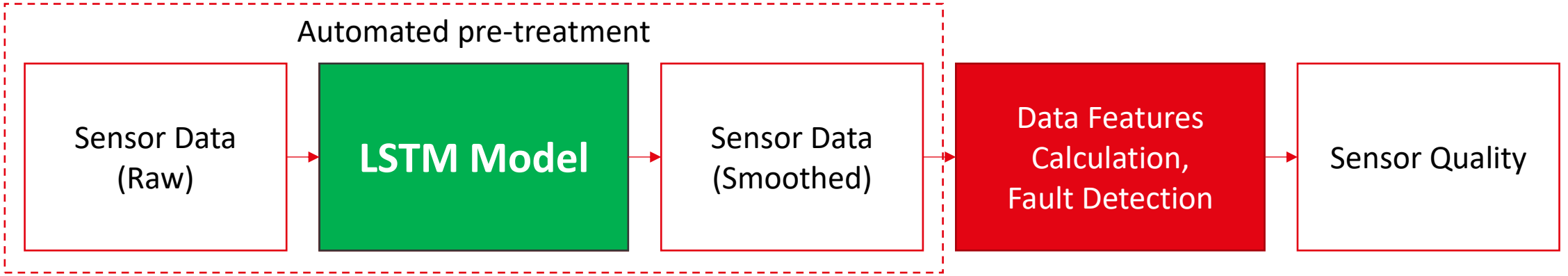


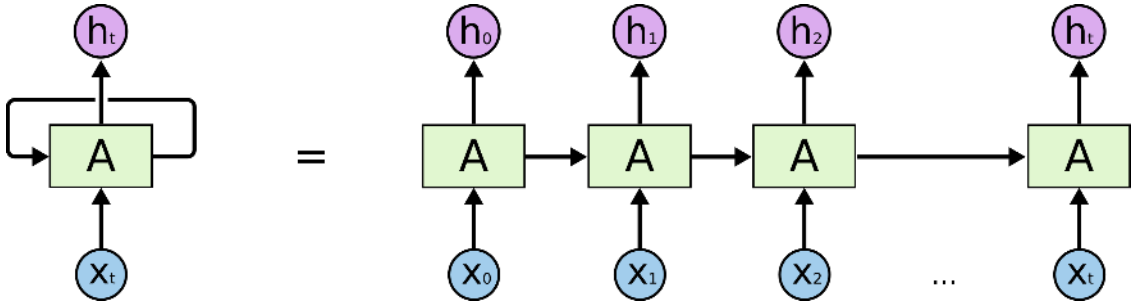
About 8% of data is considered as doubtful or not valid
(typically between 5 and 50% data loss)

UNIVERSITÉ LAVAL

Data quality control – Sensors are not perfect

- A new approach, based on LSTM methodology (Nathalia Covre)



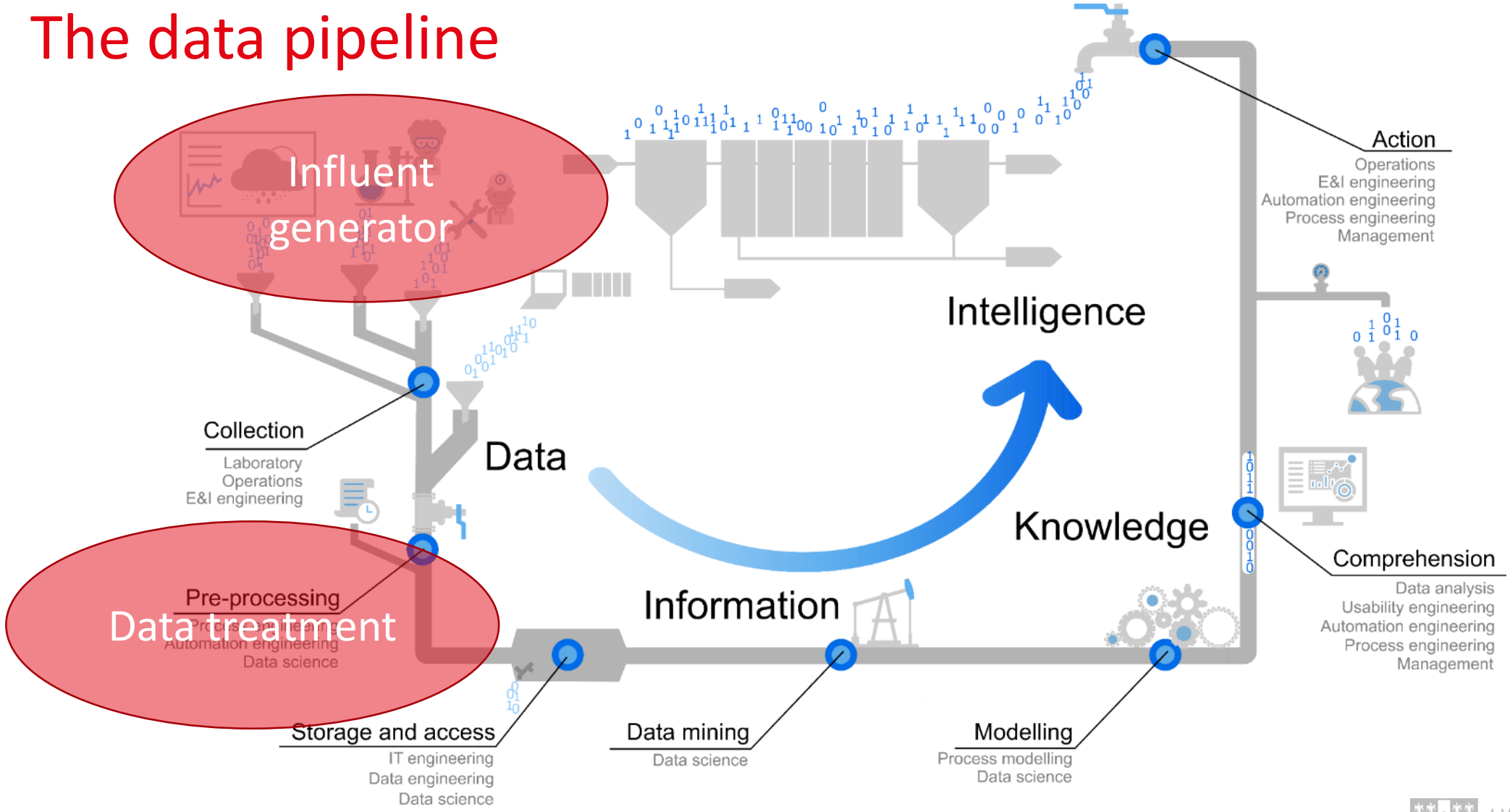


x: input data
h: predictions
A: model

LSTM : Long Short Term Memory RNN

UNIVERSITÉ LAVAL

The data pipeline



A critical review of the data pipeline: how wastewater system operation flows from data to intelligence.
Therrien J.D., Nicolai N., Vanrolleghem P. A. (2020) Water Science & Technology. DOI: 10.2166/wst.2020.393

UNIVERSITÉ LAVAL

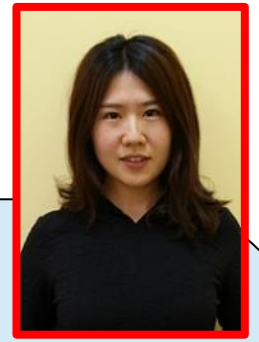
Influent generator

- Influent wastewater is main disturbance of treatment plant
  - Flow rates (rain, snowmelt, groundwater infiltration)
  - Pollutant concentrations
- Data collection is really difficult
- High-frequency long time series are needed to cover temporal variability
  - Diurnal
  - Weekly
  - Seasonal

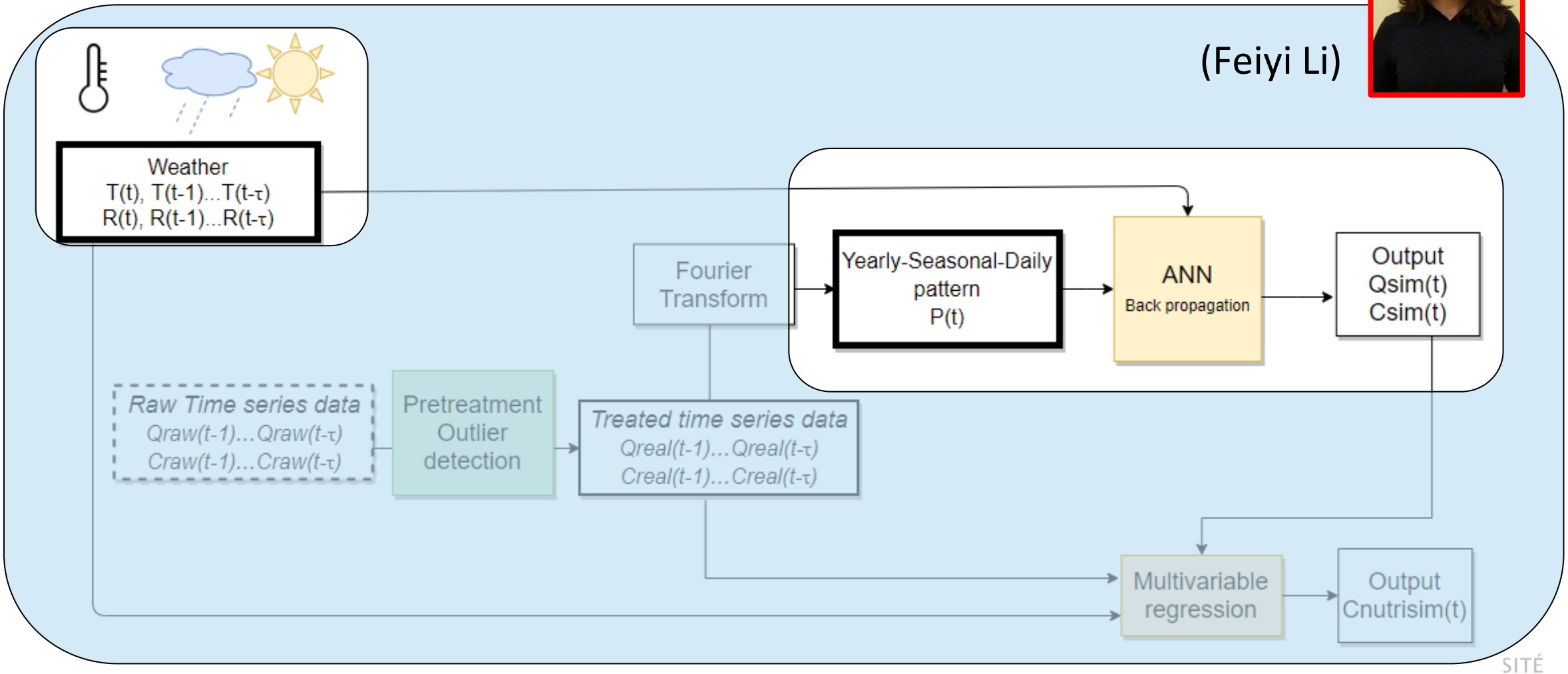
- **With limited data, generate long time series → Influent generator**

UNIVERSITÉ LAVAL

Influent generator – Methodology



(Feiyi Li)



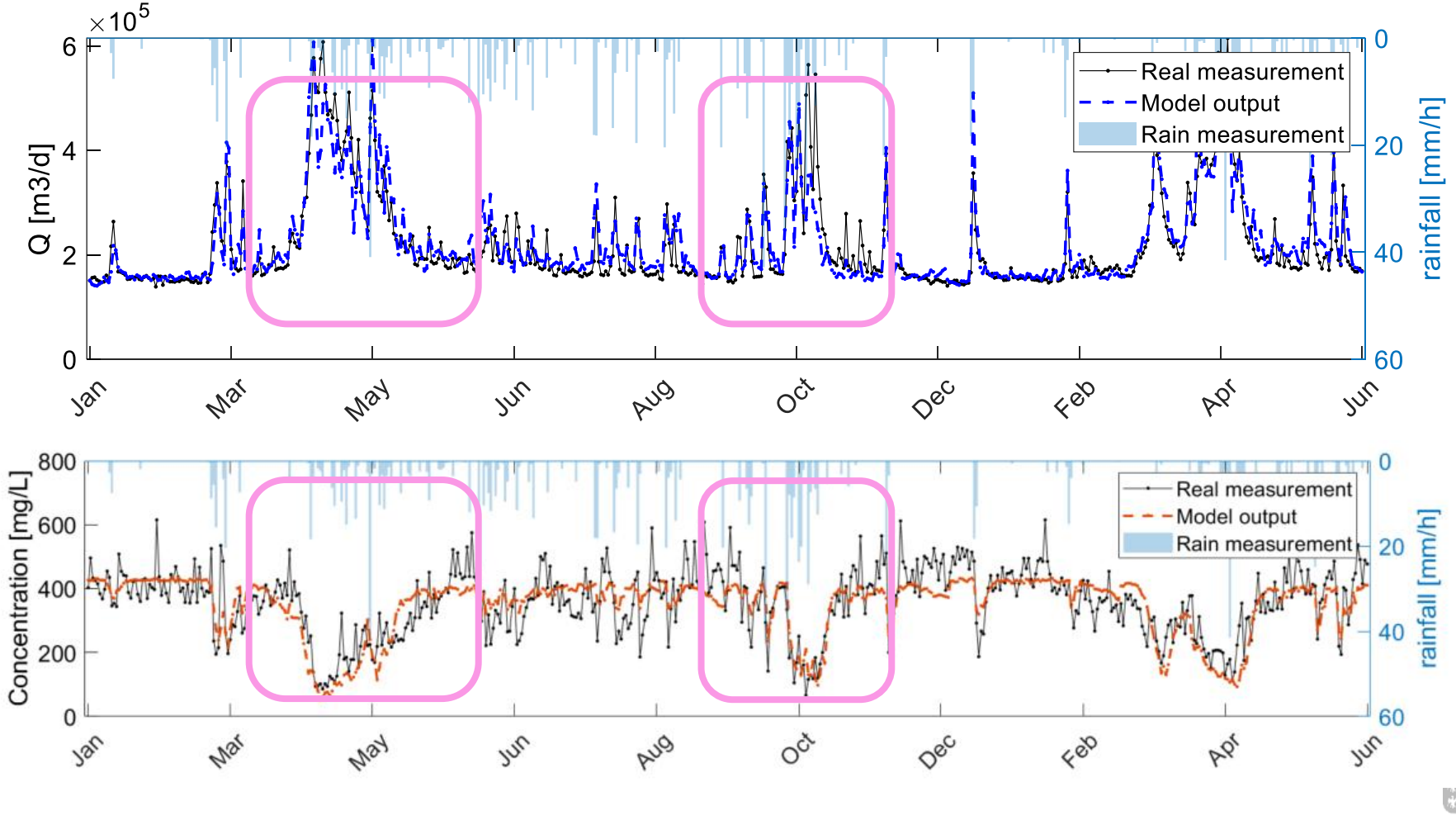
Li F. and Vanrolleghem P.A. (2022) An essential tool for WRRF modelling: A realistic and complete influent generator for flow rate and water quality based on data-driven methods. Wat. Sci. Tech., 85, 2722-2736.

SITÉ LAVAL

Influent generator – Results for Québec East plant

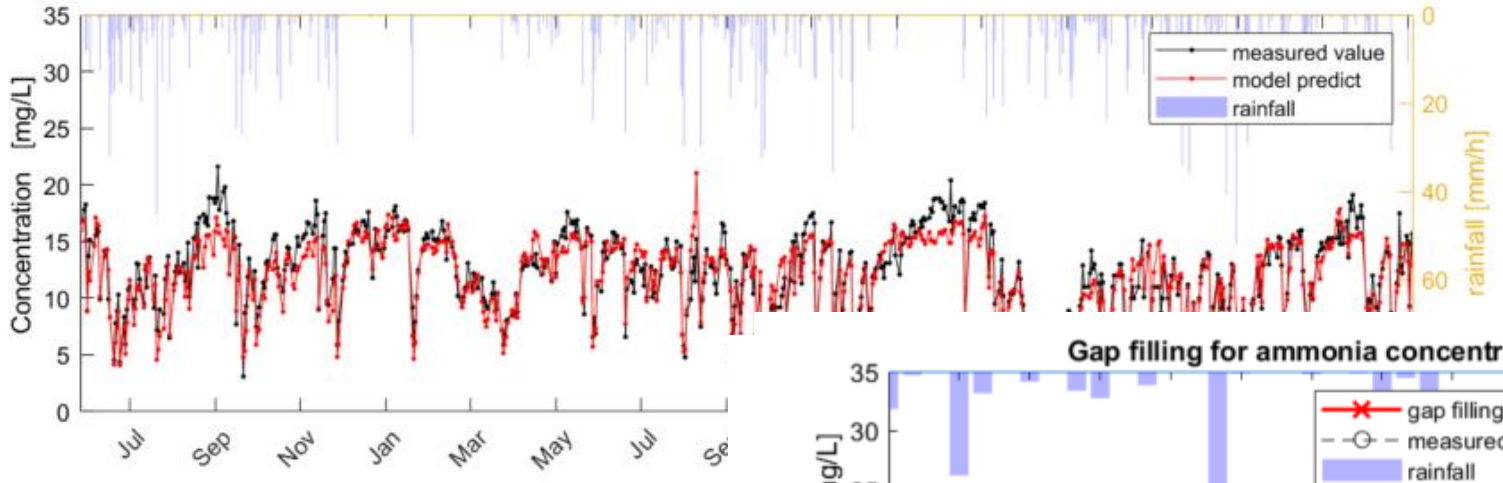
UNIVERSITÉ LAVAL

Influent generator – Results for Québec East plant

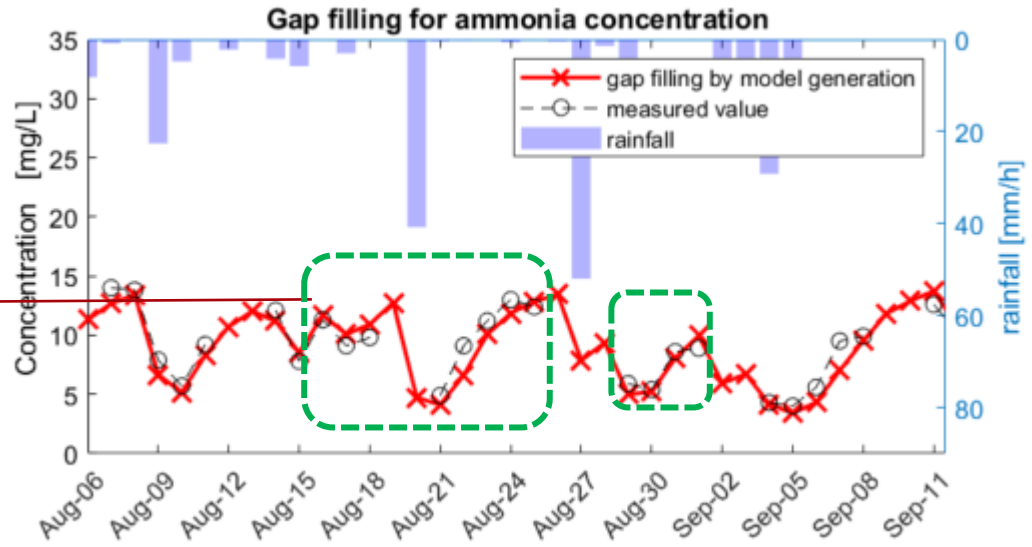


Influent generator – Gap filling for Québec East plant

- Ammonia concentration

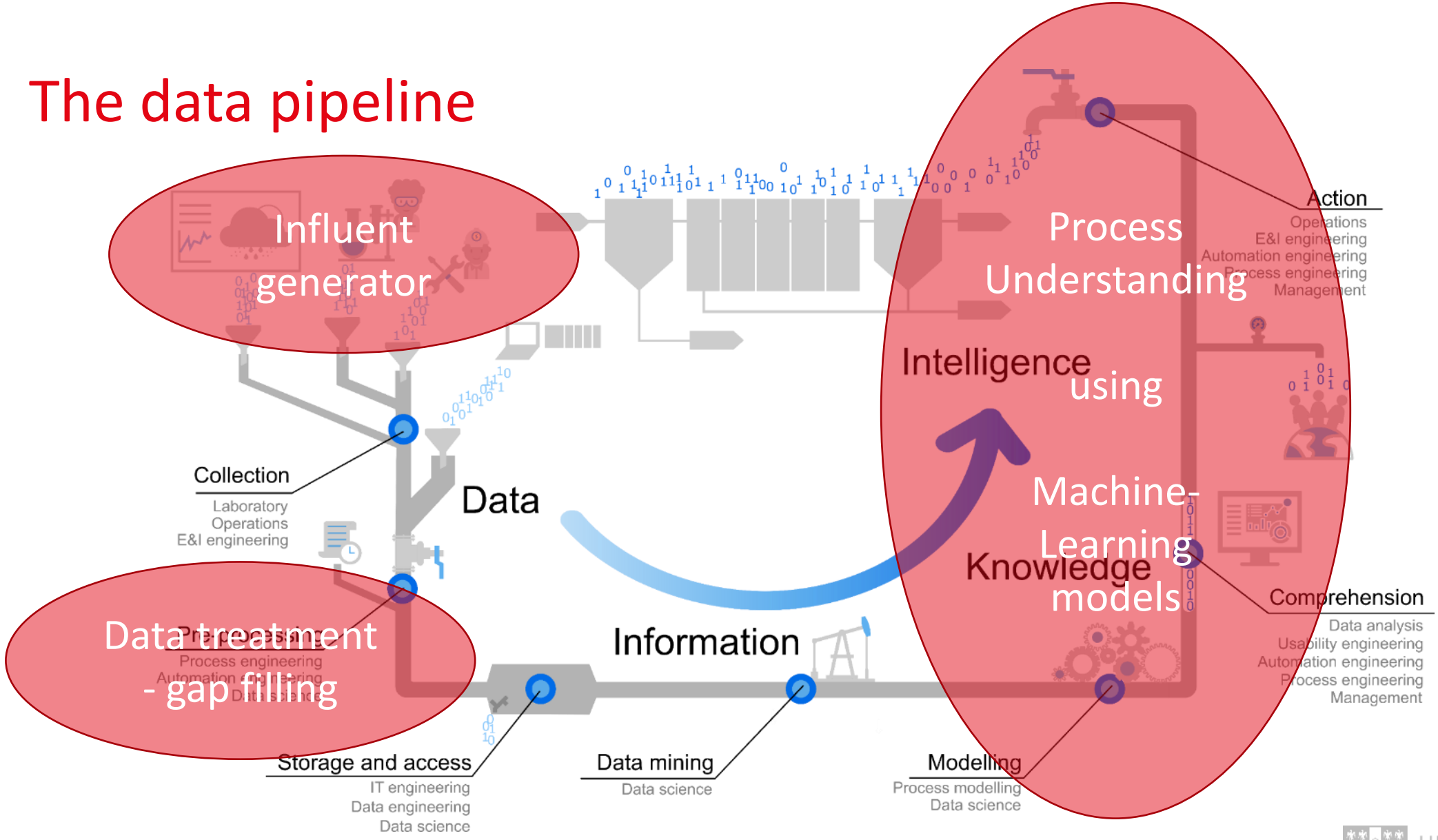


Simulated data



UNIVERSITÉ LAVAL

The data pipeline



A critical review of the data pipeline: how wastewater system operation flows from data to intelligence.
Therrien J.D., Nicolai N., Vanrolleghem P. A. (2020) Water Science & Technology. DOI: 10.2166/wst.2020.393

UNIVERSITÉ LAVAL

Process understanding – Machine “learning”

- N_2O emission modelling (Mostafa Khalil)

- Why not using mechanistic models?
  - No consensus on model selection
  - Challenging calibration and limited identifiability: affinity constants differ by up to two orders of magnitude
  - Failure to depict process dynamics in long-term full-scale datasets
  - Liquid-gas mass transfer complications

Example of mechanistic model parameters

Parameter	Min	Max
η_{AOB}	0.053	0.5
$K_{AOB\_NO_2\ NO}$	0.14	8
$K_{AOB\_NO\ NO}$	0.004	0.1
$K_{HB\_S\_NOR}$	0.56	20
$K_{HB\_i\_O_2\_NOR}$	0.067	1

- Can we use artificial intelligence/machine learning as an alternative?

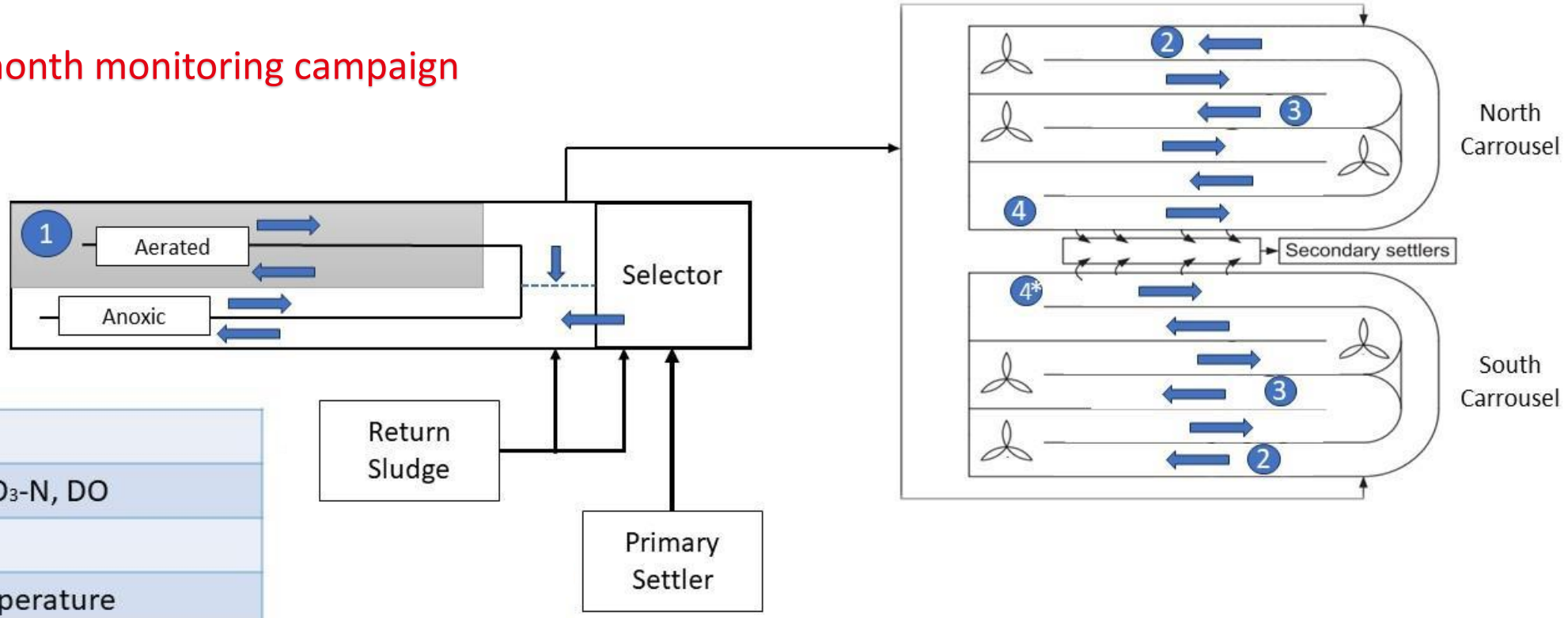
Khalil M., AlSayed A., Liu Y. and Vanrolleghem P.A. (2023) Machine learning for modeling N2O emissions from wastewater treatment plants: Aligning model performance, complexity, and interpretability. Wat. Res., 245, 120667.

UNIVERSITÉ LAVAL

Process understanding – Machine “learning”

- N_2O emission modelling

A 16-month monitoring campaign



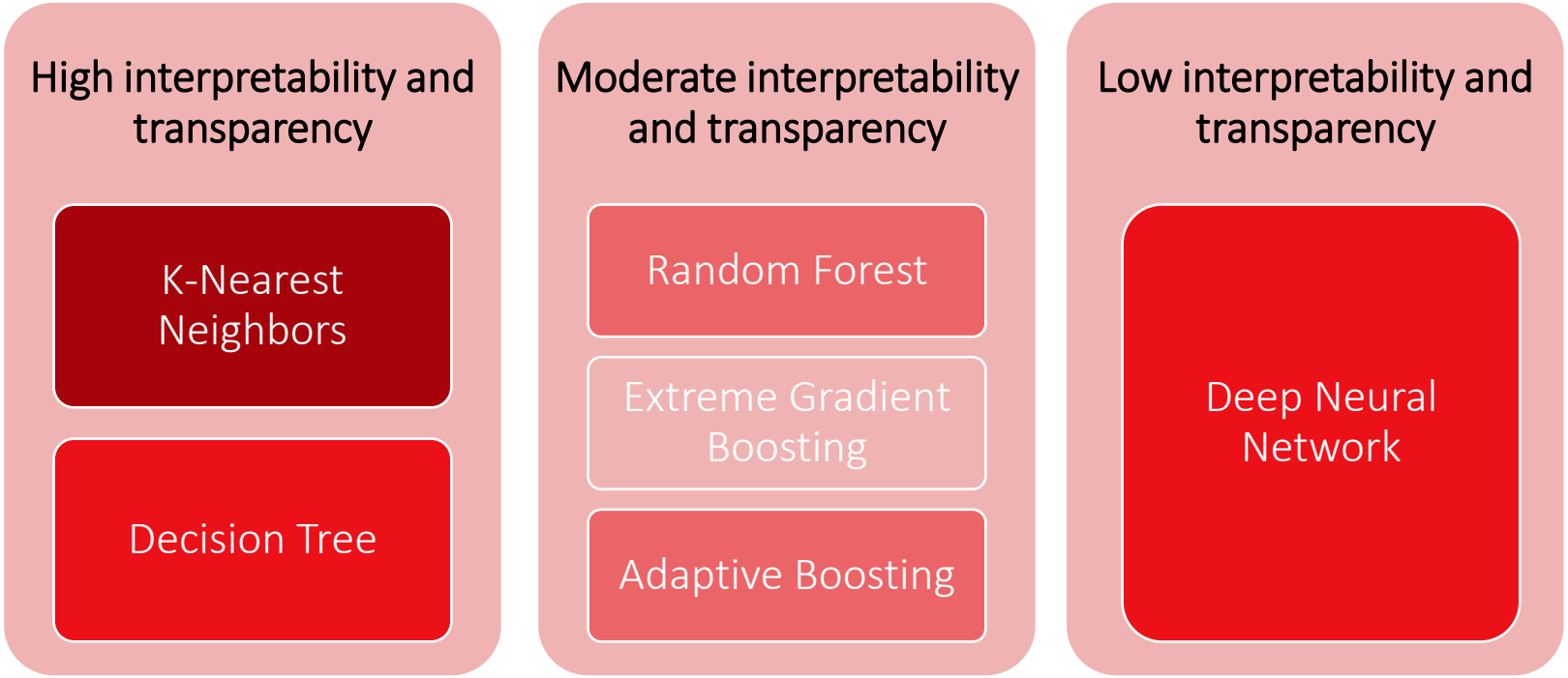
Location	Probe
1	NH_4-N, NO_3-N, DO
2	DO1, TSS
3	DO2, Temperature
4	DO3, NH_4-N, NO_3-N,NO_2-N
4*	DO3, NH_4-N, NO_3-N

Layout of Kralingseveer WWTP (Daelman et al., 2015)

UNIVERSITÉ LAVAL

Process understanding – Machine “learning”

- ML-model selection

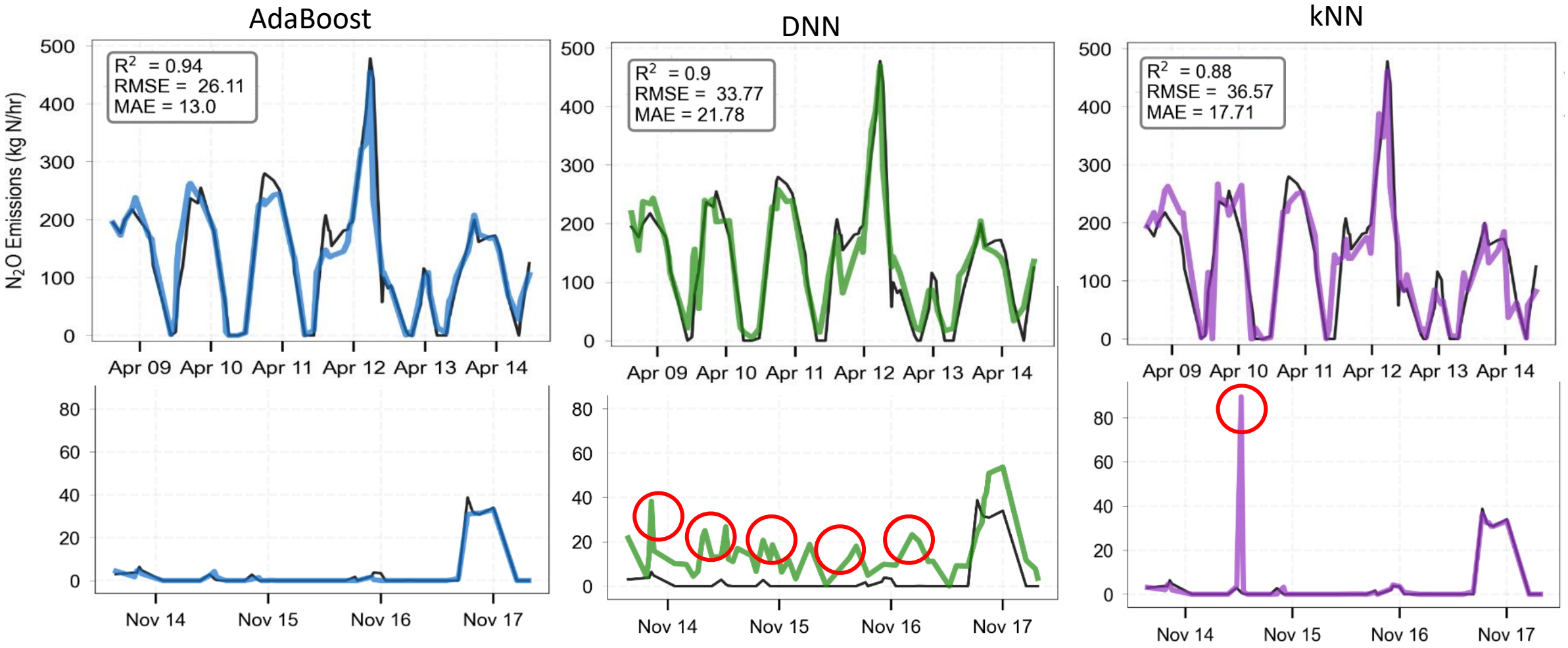


- **Interpretability**: Degree to which a model's decision-making process can be understood by humans
- **Transparency**: Clarity of a model's architecture, parameters, and training process

UNIVERSITÉ LAVAL

Process understanding – Machine “learning”

- ML model testing performances → Simple (AdaBoost, kNN) works!

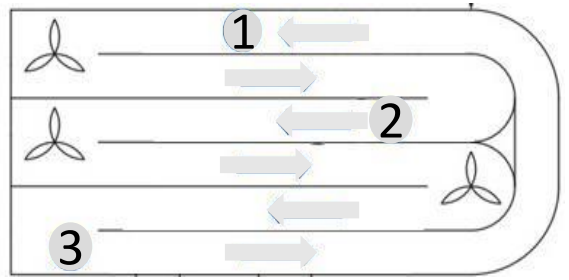


Process understanding – Machine “learning”

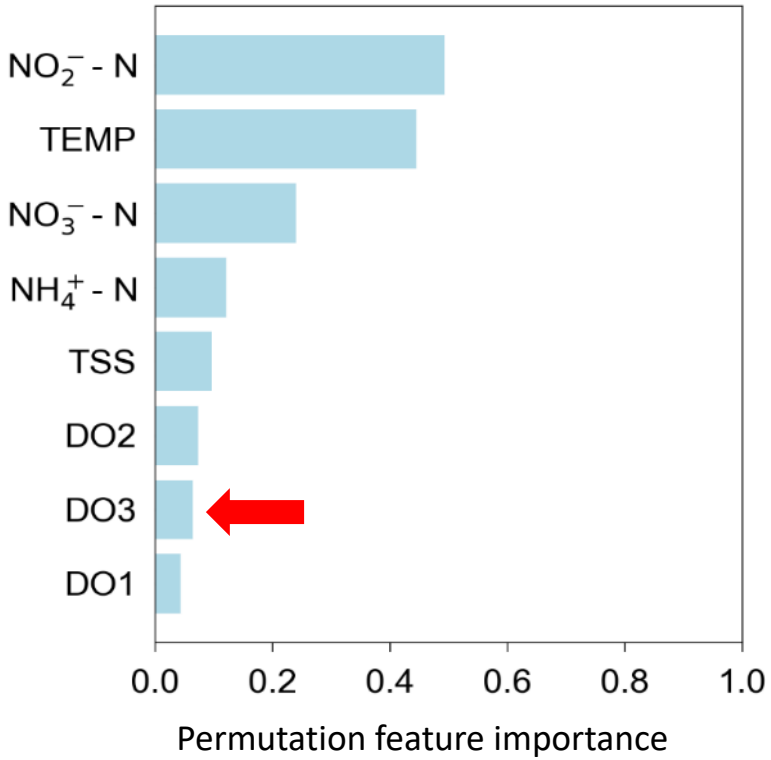
- ML as a method to learn

What features are the most important for the model to make N_2O-prediction?

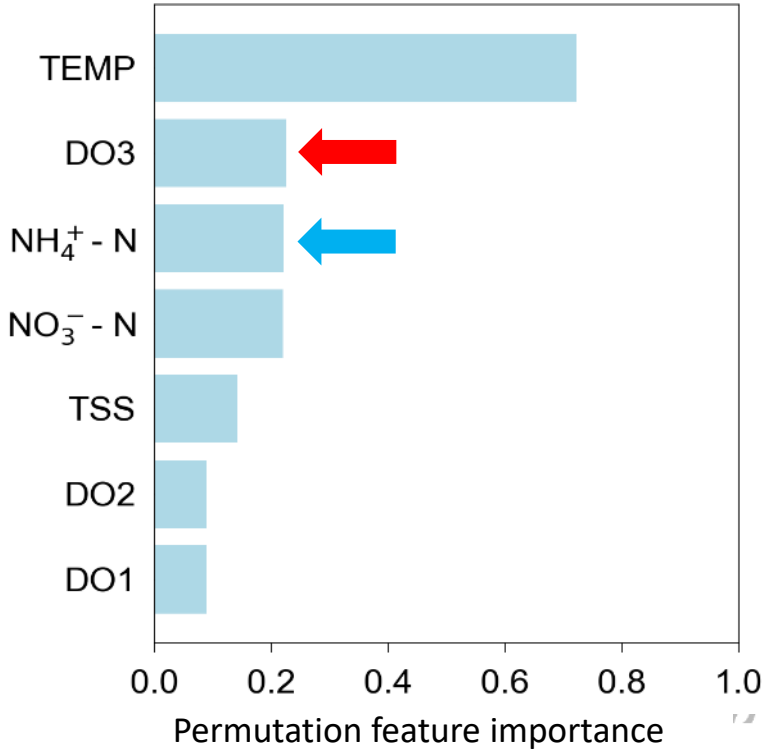
DO Measurement Locations



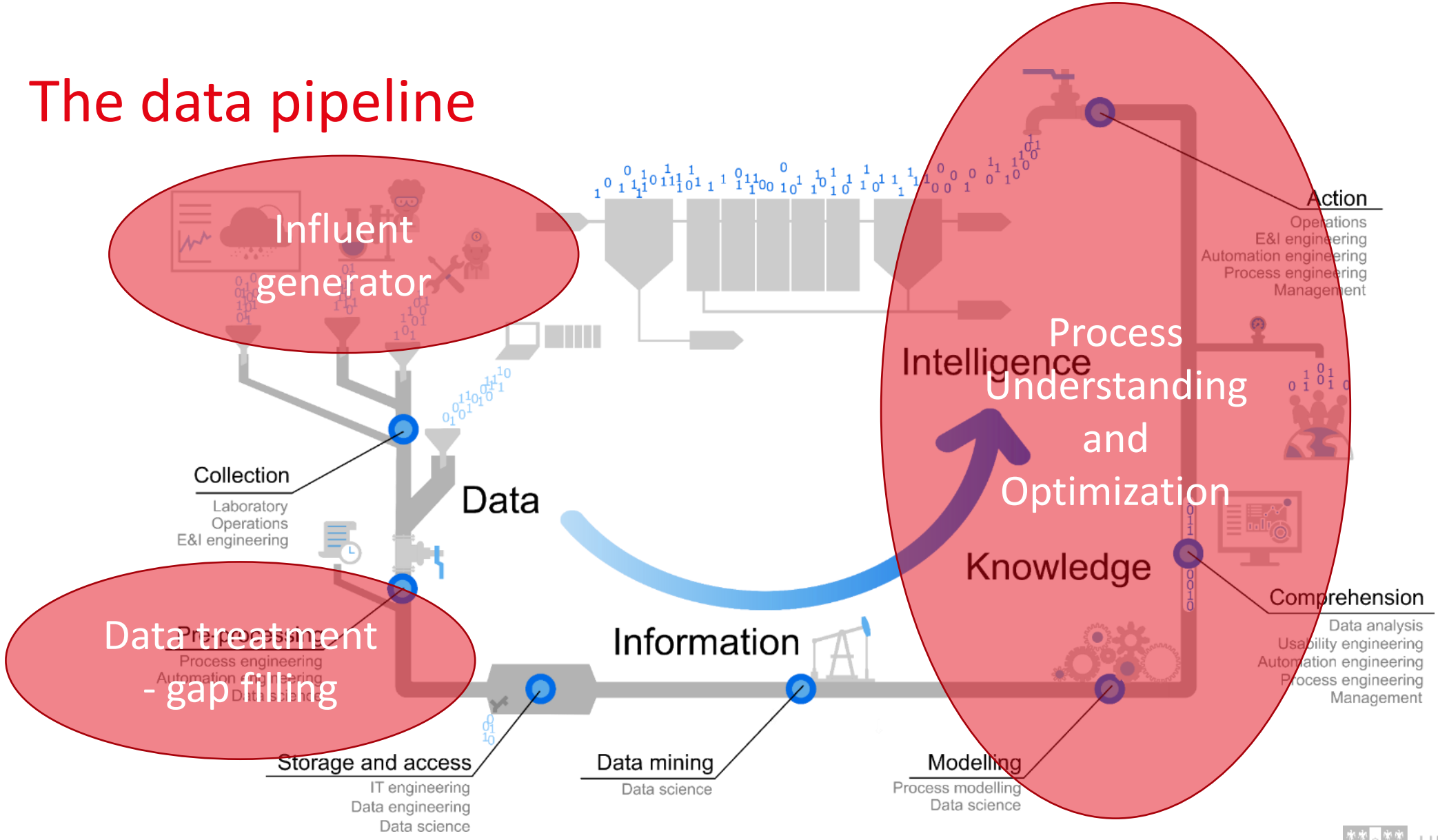
Same NO_2 sensor location



Remove NO_2 from input features



The data pipeline



A critical review of the data pipeline: how wastewater system operation flows from data to intelligence.
Therrien J.D., Nicolai N., Vanrolleghem P. A. (2020) Water Science & Technology. DOI: 10.2166/wst.2020.393

UNIVERSITÉ LAVAL

Digital transformation of the water sector

- It is happening now (finally!)
- Challenges are not small
- Data collection remains difficult
- AI and ML tools have become sufficiently mature to be used in many practical situations to great benefit

- Training of water professionals will be key

- Under-promise – Over-deliver

Vanrolleghem P.A., Khalil M., Serrao M., Sparks J. and Therrien J.D. (2025)
Machine learning in wastewater: Opportunities and challenges—"**not everything is a nail!**".
Current Opinion in Biotechnology, 93, p.103271.

the international water association

Machine learning supporting the operation of wastewater treatment processes:

Get ready – the future is near!

PETER A. VANROLLEGHEM

NATHALIA COVRE, MOSTAFA KHALIL, FEIYI LI, BERNARD PATRY, MARCELLO SERRAO, JEFF SPARKS, JEAN-DAVID THERRIEN

inspiring change

UNIVERSITÉ LAVAL

PRESENTATION 2

AI-Powered Asset Management: Turning Your Asset Plan into a Living Tool

Nicolas St-Gelais
Scientific Director
CANN Forecast, Canada

Nicolas St-Gelais is the Scientific Director at CANN Forecast, a company specializing in the development of artificial intelligence tools for proactive water management.

During his postdoctoral research, he studied the impacts of climate change on lake water quality using advanced modelling approaches.

In 2016, he participated in the AquaHacking competition and won with the solution InteliSwim. This achievement led to the creation of CANN Forecast, which now supports more than 50 municipalities across Canada, France, and the United States in improving water infrastructure management through data and AI.

IWA Webinar

Auto-AMP: AI for water Infrastructure Management

CANN FORECAST | Smart Water Management

ABOUT CANN FORECAST

CANN Forecast AI Journey

1

2017 — Water Quality

Developed **InteliSwim**, a ML model to predict water quality — built during a Hackathon

2

2018 — Water Main Breaks

Developed **InteliPipes**, a ML model to predict water main breaks for the City of Montreal

3

2021 — Flow Forecasting

Developed **InteliFlow**, a ML model to predict risk of overflow with the Peel Region

4

2023 — Asset Management

Developed **Auto-AMP**, an automated tool to generate asset management plans using InteliPipes and LLM

5

2024 — Asset Chatbot

Launched conversational AI for infrastructure decision support

ABOUT CANN FORECAST

Naysan Saran
Co-founder & CTO
Applied mathematics and AI systems design

Nicolas St-Gelais, Ph.D.
Co-founder & Scientific Director
Expertise in data science and aquatic systems modeling

Jennifer Cai, M. Sc.
Data Scientist
Time series forecasting & AI-driven quality control

Anush Chaturvedi
Front-end Developer

Chris Pisaric
Senior Water Systems Consultant
30+ years of operational experience

Julien Magne
Frontend GIS Developer
Geospatial systems and visualization

Cooper Albano, M. Sc.
Data Scientist
Urban Hydrology Specialist

Sara Zapata-Marin, Ph. D.
Statistician
Bayesian statistics & causal inference

Andre Della Libera Zanchetta, PhD
Full-Stack Developer
Hydrological systems & software engineering

Rebecca Dziedzic, PhD
Infrastructure Consultant
Professor in Civil, environmental and building engineering

ABOUT CANN FORECAST

Data-Driven Innovation in Water & Wastewater

50+ Utilities

Trusted by water and wastewater utilities across North America

Co-Developed Solutions

Every product is built alongside clients — ensuring real-world applicability

Peer-Reviewed Science

Published research in leading journals validates our methodology

ABOUT CANN FORECAST

CANN Forecast AI Journey

1

2017 — Water Quality

Developed **InteliSwim**, a ML model to predict water quality — built during a Hackathon

2

2018 — Water Main Breaks

Developed **InteliPipes**, a ML model to predict water main breaks for the City of Montreal

3

2021 — Flow Forecasting

Developed **InteliFlow**, a ML model to predict risk of overflow with the Peel Region

4

2023 — Asset Management

Developed **Auto-AMP**, an automated tool to generate asset management plans using InteliPipes and LLM

5

2024 — Asset Chatbot

Launched conversational AI for infrastructure decision support

850 water mains break in North America every single day

InteliPipes

For a Medium Utility

500

Annual Breaks

Water main failures per year at a typical mid-size utility

$5M

Direct Repair Cost

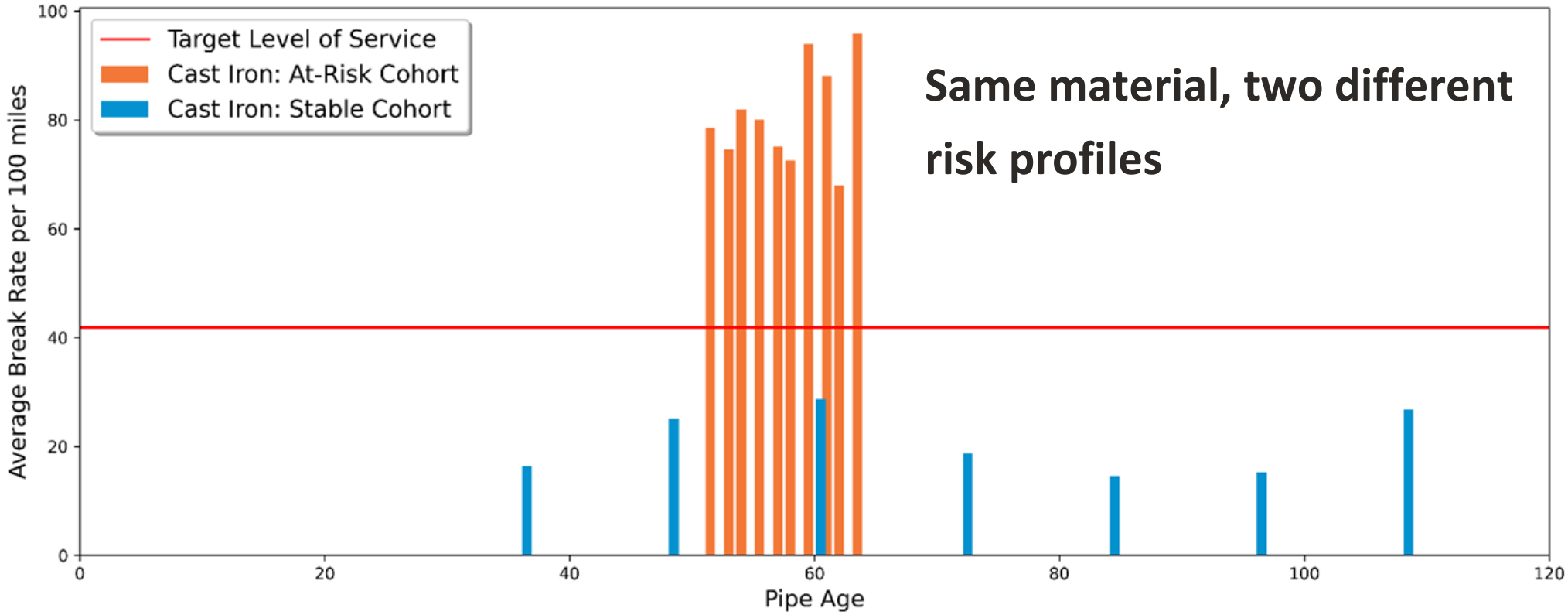
In direct repair expenditures — not counting indirect disruption costs

25%

Water Lost

Of treated drinking water lost to leaks and main breaks every year

The Problem with age-based replacement



InteliPipes for Water

Structural Factors

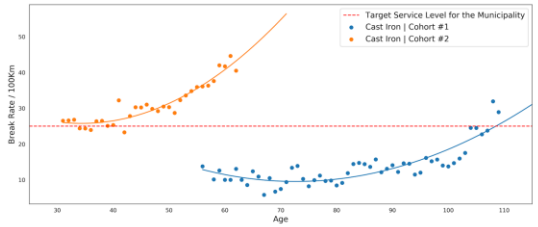
installation date, age, diameter, length, break history, pipe location

Operational & Environmental Factors

pressure average and fluctuation, seasonal patterns of temperature, frost & thaw, Pipe lining, soil type, road type, traffic loads, population density, C factor, breaks in vicinity of pipe

Machine Learning

Pipe Cohorts, sorted by level of risk

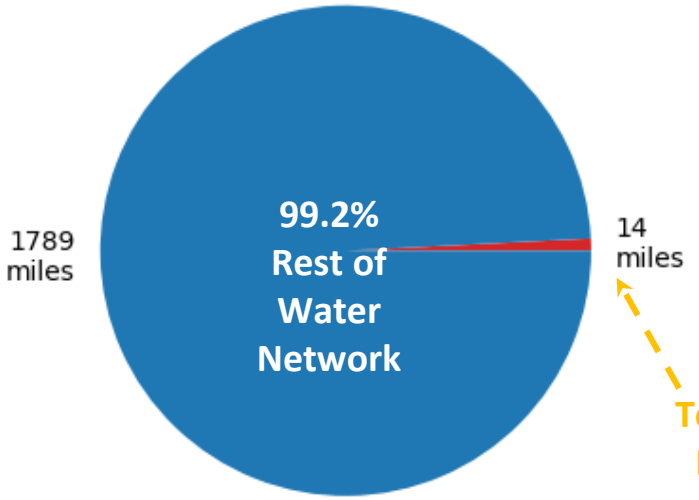


Degradation curves per cohort

Map of your most at-risk areas

InteliPipes

In 2020, the InteliPipes AI model identified 14 miles of pipes that have highest probability of failure



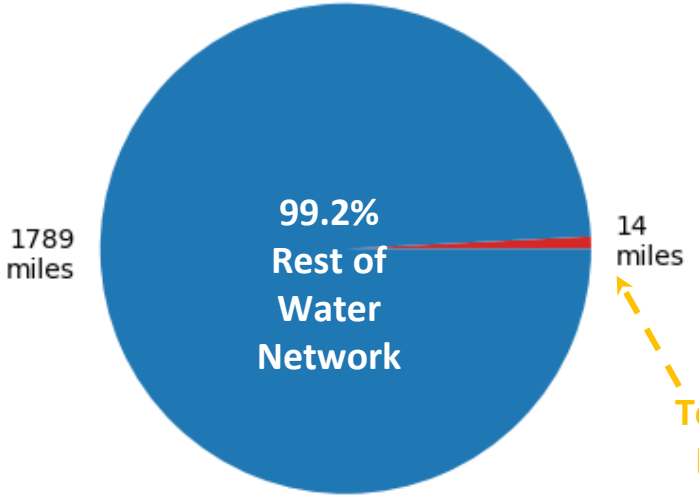
We then waited two years for breaks to accumulate

...

Less than 1% of total network identified by InteliPipes using data from 2015 to 2020 - was responsible for **14% of breaks in 2021** and **22% of breaks in 2022**

InteliPipes

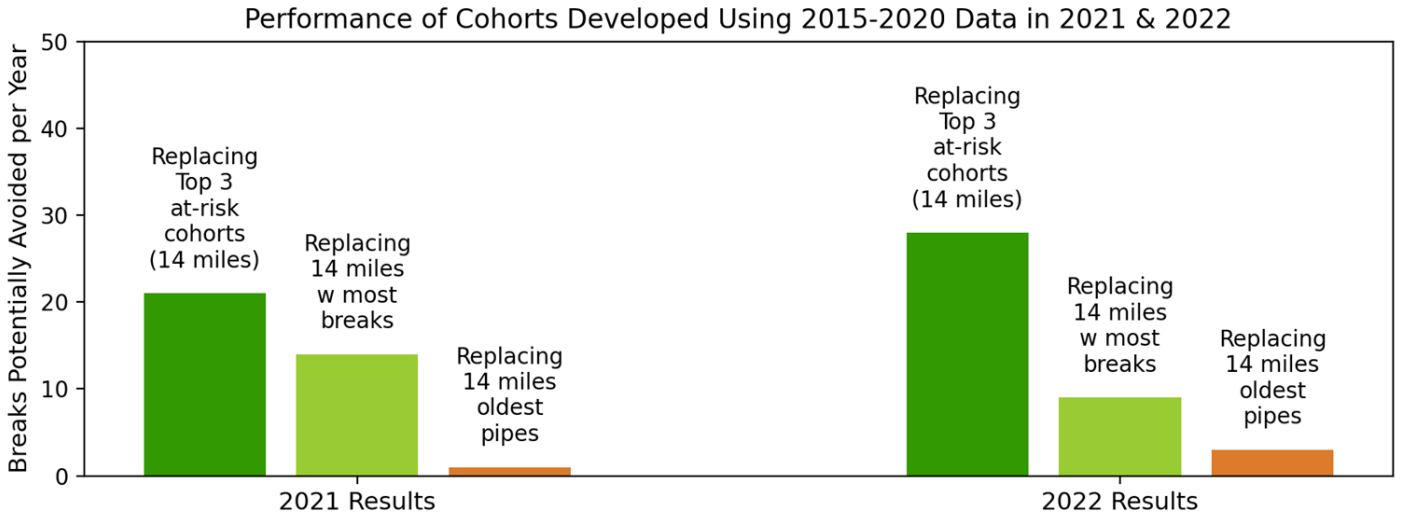
In 2020, the InteliPipes AI model identified 14 miles of pipes that have highest probability of failure



Top 3 Most At-Risk Cohorts

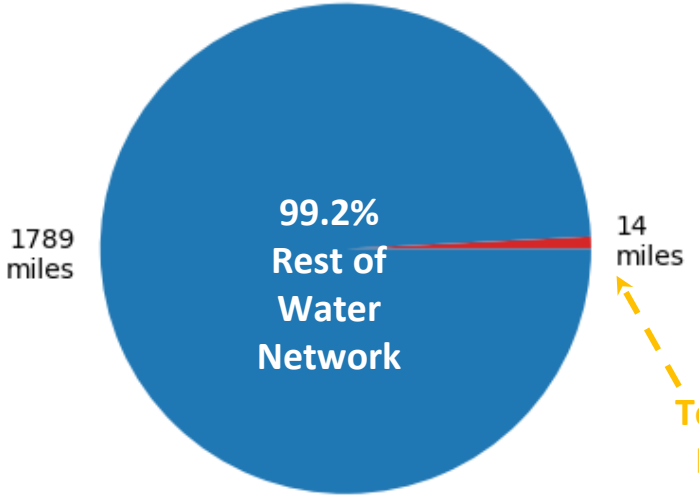
We then waited two years for breaks to accumulate …

Less than 1% of total network identified by InteliPipes using data from 2015 to 2020 - was responsible for **14% of breaks in 2021** and **22% of breaks in 2022**



InteliPipes

In 2020, the InteliPipes AI model identified 14 miles of pipes that have highest probability of failure



We then waited two years for breaks to accumulate …

Less than 1% of total network identified by InteliPipes using data from 2015 to 2020 - was responsible for **14% of breaks in 2021** and **22% of breaks in 2022**

Performance on Future Breaks (2021-2022)

Age-Based Replacement: 1 break prevented in 2021 and 3 breaks prevented in 2022

Cumulative Worst Breaks Based Replacement: 14 breaks in 2021 and 9 breaks in 2022

InteliPipes AI Model Based Replacement: 21 breaks in 2021 and 28 breaks in 2022

A MEASURABLE IMPACT

InteliPipes for Drinking Water

Results of pilot project with Peel,

The environmental benefits

Up to 12x more accurate than the age-based models, which translates into:

- 14,000 cubic meters of water saved every year, per 100Km of network
- Average energy saved: 8,680 kWh/year, per 100Km of network

The financial benefits

- Emergency repair savings: USD$ 54,000/year, per 100Km of network
- Indirect savings due to the consequences of water main breaks: On average, USD$ 216,000/year, per 100Km of network

InteliPipes for Urban Flooding

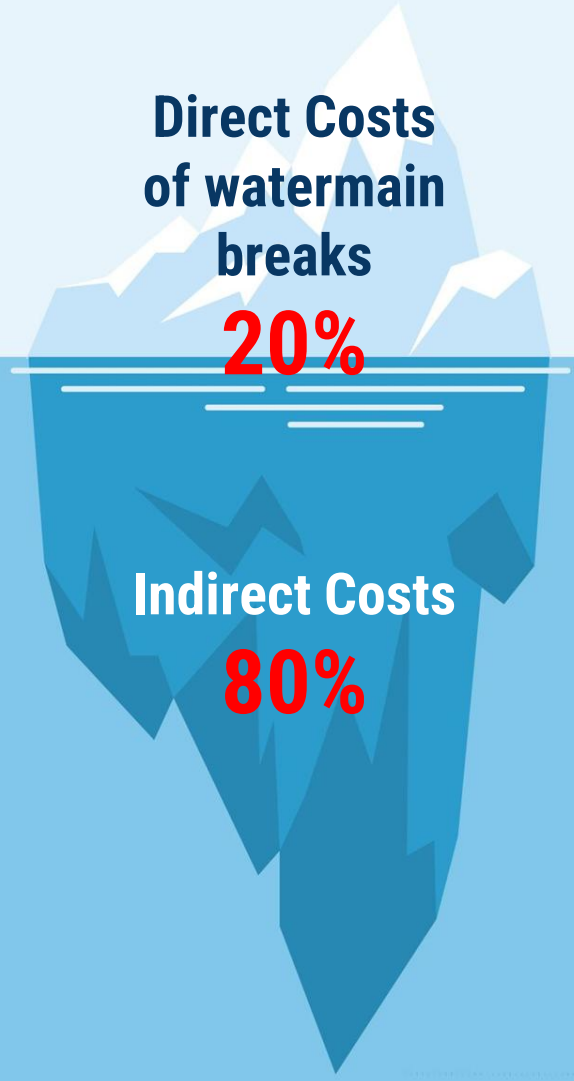
Pilot project with Peel Region, Ontario, Canada

Operational benefits

- Our tool predicts wastewater flow in under a minute, a significant improvement over the traditional models which typically take an hour.
- Our predictions for storm events exceeding 15mm demonstrated greater accuracy than those provided by the InfoWorks hydraulic model

Projected financial benefits

- Our model is significantly more affordable than traditional alternatives, costing only about 25-50% of what those models typically do.



ABOUT CANN FORECAST

CANN Forecast AI Journey

1

2017 — Water Quality

Developed **InteliSwim**, a ML model to predict water quality — built during a Hackathon

2

2018 — Water Main Breaks

Developed **InteliPipes**, a ML model to predict water main breaks for the City of Montreal

3

2021 — Flow Forecasting

Developed **InteliFlow**, a ML model to predict risk of overflow with the Peel Region

4

2023 — Asset Management

Developed **Auto-AMP**, an automated tool to generate asset management plans using InteliPipes and LLM

5

2024 — Asset Chatbot

Launched conversational AI for infrastructure decision support

Asset Management Plans Are Static — But Reality Is Not

Current Reality:

- Asset management plans are compliance-driven
- Updated once every few years
- Based on fragmented and outdated data
- Limited connection to real operations

The Gap:

- Decisions are not aligned with real-time conditions
- Difficult to assess financial impact of actions
- Limited visibility on level of service impacts
- Plans become static documents instead of decision tools

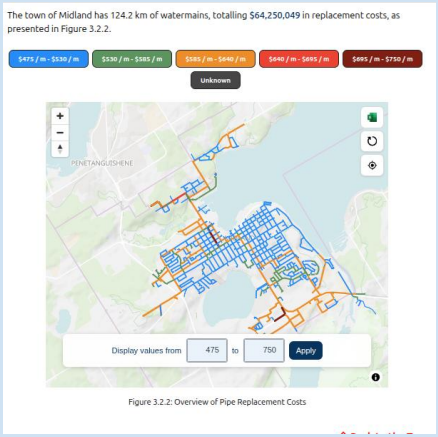
Utilities need dynamic, data-driven asset management — not static reports

Asset management

Typical Client:

Muncipalities & utilities (e.g. engineer, asset manager)

Dynamic asset management plan



Clean structured data is stored in secured centralized database

Uploads data to the platform:

- Asset information (e.g. inspections, age, …)
- reports (excel, pdfs)
- etc…

Configure their needs

- Available budgets
- Level of service
- Climate change strategy

AI tool box

Clean, structure and enrich data automatically

LLM

ETL

INTELIPIPES

Inside the AI Tool Box

ETL Pipeline:

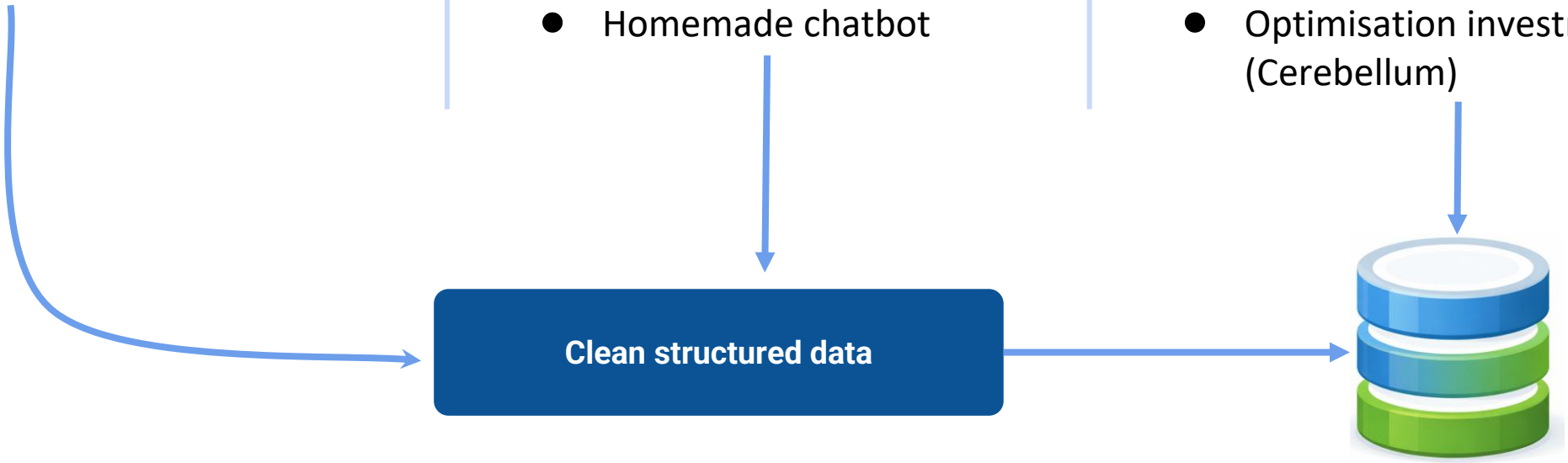
- Clean and validate data
- Detect anomalies

Private LLM deployment :

- Extract info from reports
- Automated text creation
- Homemade chatbot

CANN's AI models:

- ML models to identify most at risk assets (e.g. InteliPipes)
- Optimisation investment tool (Cerebellum)



Cerebellum: Asset optimization planing

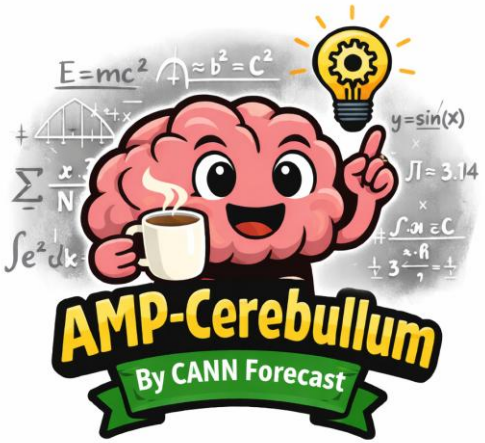
- Utilities operate under limited budgets while maintaining service levels
- Which assets should be prioritized

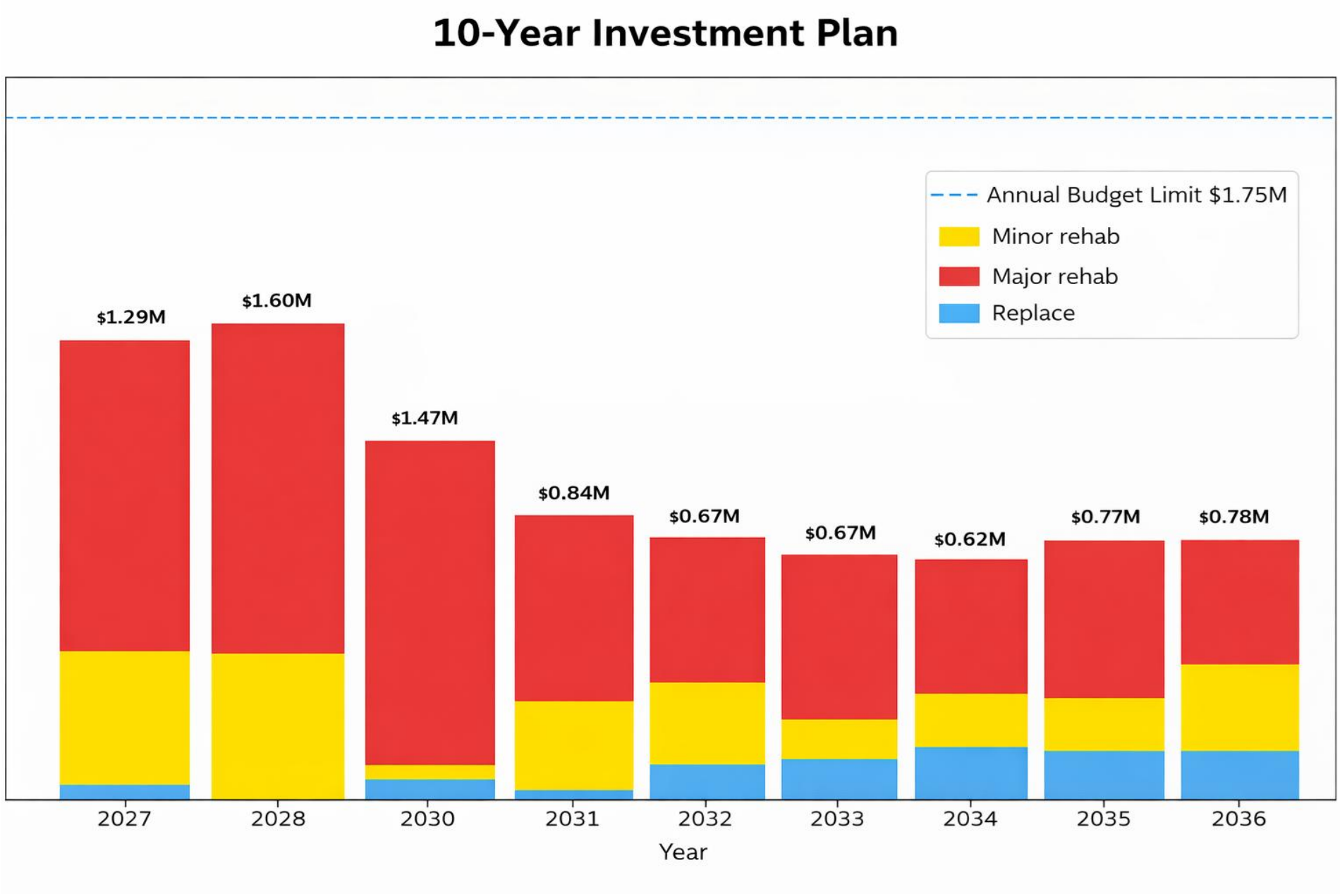
Activity	Cost	Impact on useful life
Replacement	40 000$	+ 80 years
Relining	15 000$	+ 25 years

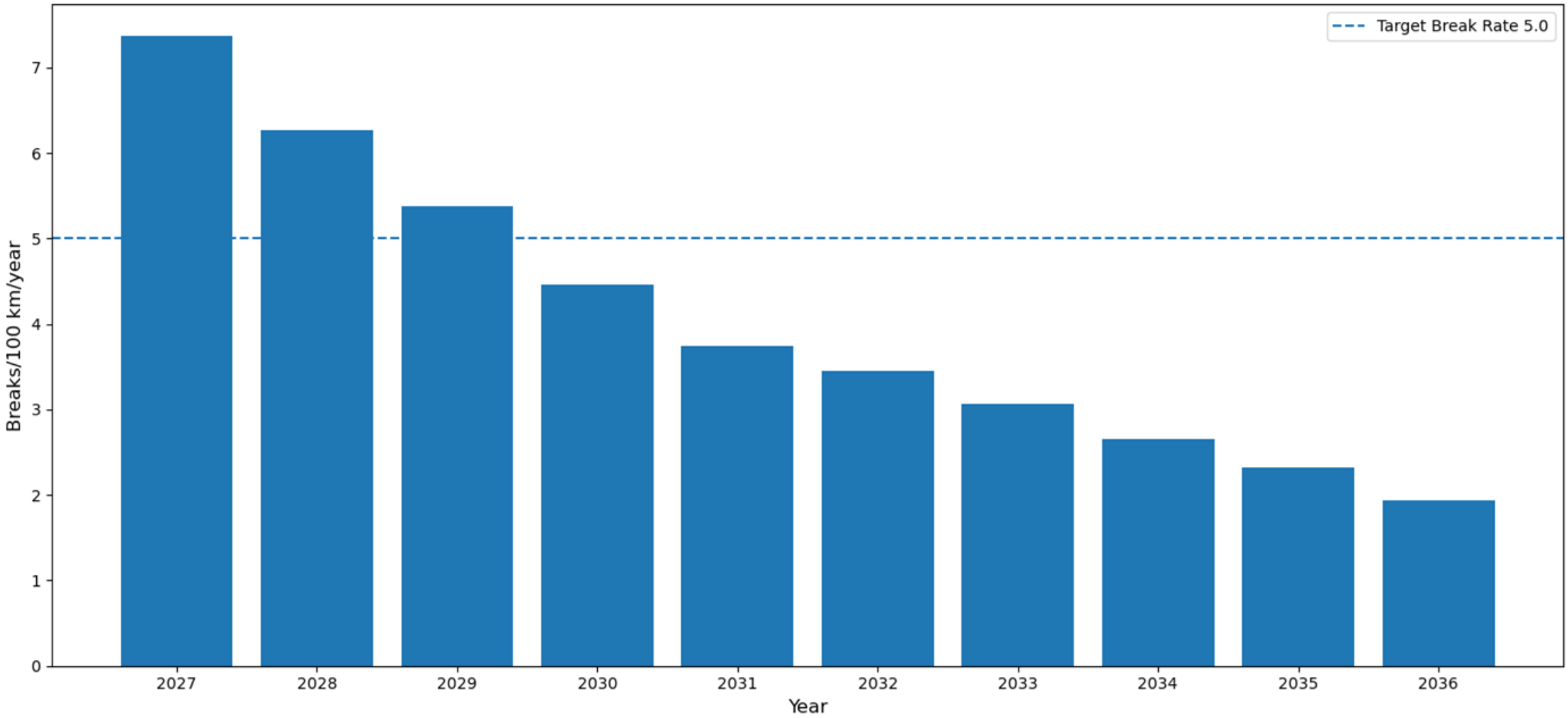
Cerebellum: Asset optimization planing

Cerebellum is an optimization tool:

- Cerebellum is an optimization tool: Learns how assets deteriorate to predict future condition and failure
- **Two modes:**
  - **Break-rate model (linear assets)** — Achieve target network break rate optimizing budged
  - **Condition (RUL) model (linear and vertical assets)** — Plans so that only up to a target share of the network stays in poor / very poor condition.
- **What you get:**
  - 10-year investment plan
  - Which assets to treat
  - When to act
  - Cost per year




10-Year Investment Plan
Annual Budget Limit $1.75M
Minor rehab
Major rehab
Replace
$1.29M
$1.60M
$1.47M
$0.84M
$0.67M
$0.67M
$0.62M
$0.77M
$0.78M
2027
2028
2030
2031
2032
2033
2034
2035
2036
Year

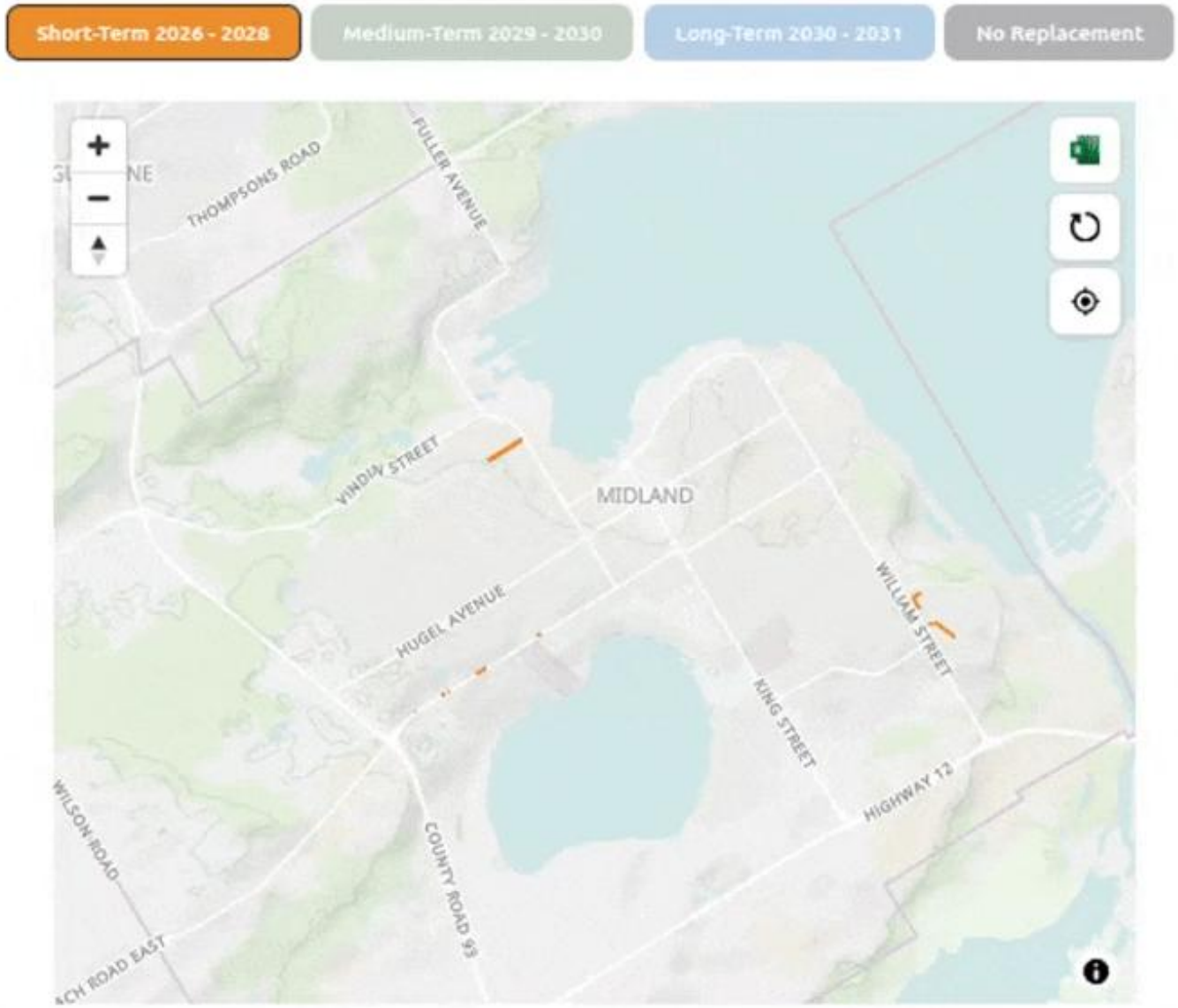

Target Break Rate 5.0
Breaks/100 km/year
0
1
2
3
4
5
6
7
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
Year

Asset management

Ce

Cere

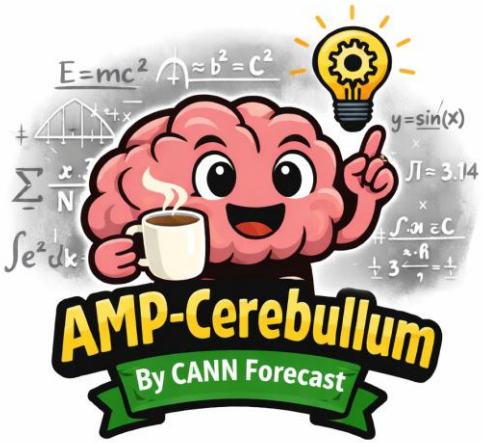
- C
  c
- T
- V

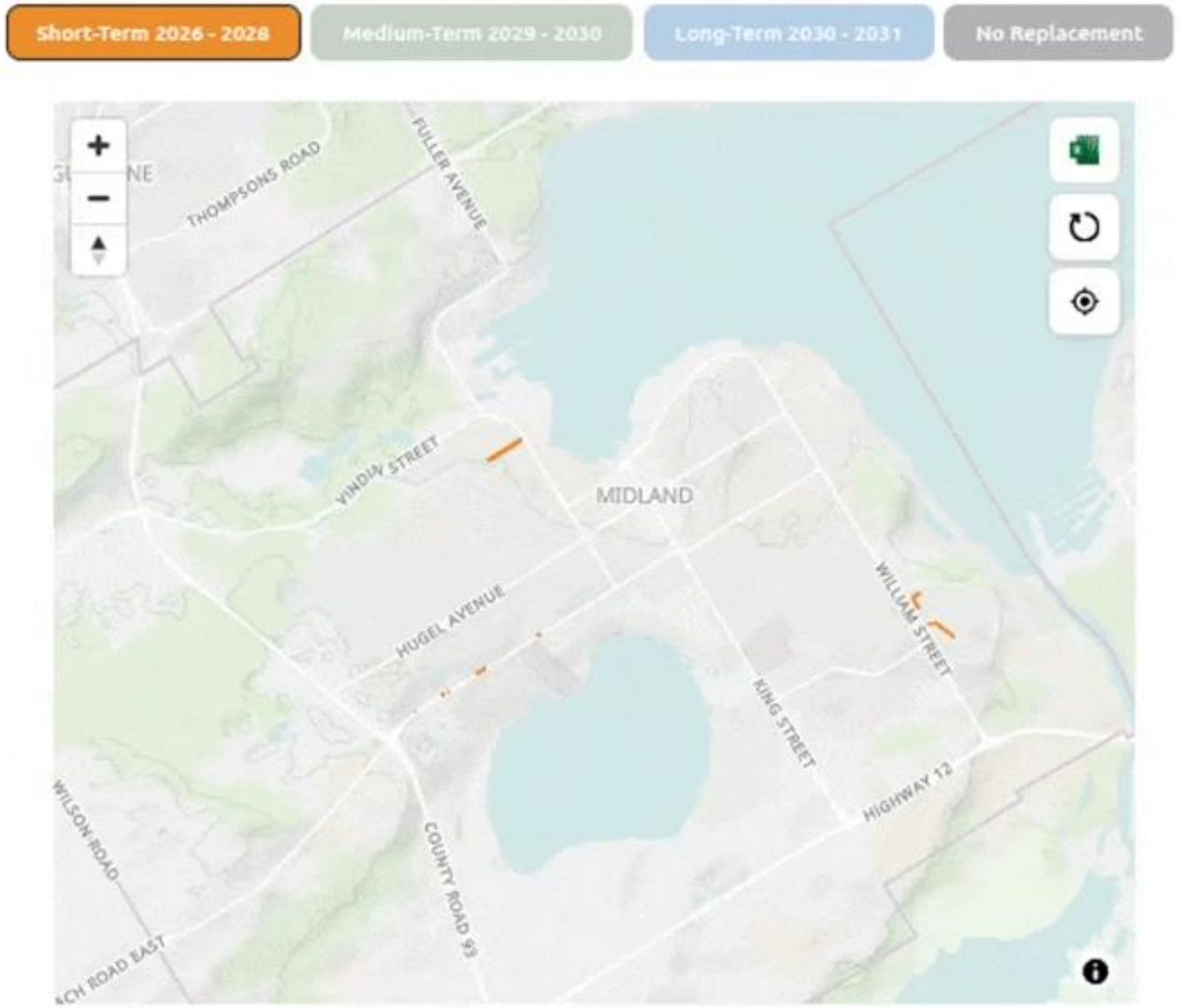


edict future

ate

only up to a





Now scale this across all asset types

ABOUT CANN FORECAST

CANN Forecast AI Journey

1

2017 — Water Quality

Developed **InteliSwim**, a ML model to predict water quality — built during a Hackathon

2

2018 — Water Main Breaks

Developed **InteliPipes**, a ML model to predict water main breaks for the City of Montreal

3

2021 — Flow Forecasting

Developed **InteliFlow**, a ML model to predict risk of overflow with the Peel Region

4

2023 — Asset Management

Developed **Auto-AMP**, an automated tool to generate asset management plans using InteliPipes and LLM

5

2024 — Asset Chatbot

Launched conversational AI for infrastructure decision support

Chatbot

Unlock the Hidden Value of Asset Data

- Utilities sit on a **gold mine of data**:
  - Structured (GIS, inspections, sensors)
  - Unstructured (reports, PDFs, procedures)
- This data is **hard to access** and **underused**
- **Traditional LLMs**: hallucinations, low reliability on technical data, not adapted to utility workflows

AI-powered assistant designed for utilities:

- Extracts **insights from all data sources**
- Answers **operational & asset questions**
- **Objective:** Provide **95+% accuracy when answering verifiable questions about client datasets** (Validated on a dataset of 10 000 questions).

Chatbot

Two Parallel Approaches

Structured Data Retrieval

- Provide data *structure* to the LLM
- Local tools enables complex analysis, plotting, and mapping

Consistently accurate answers about physical asset data

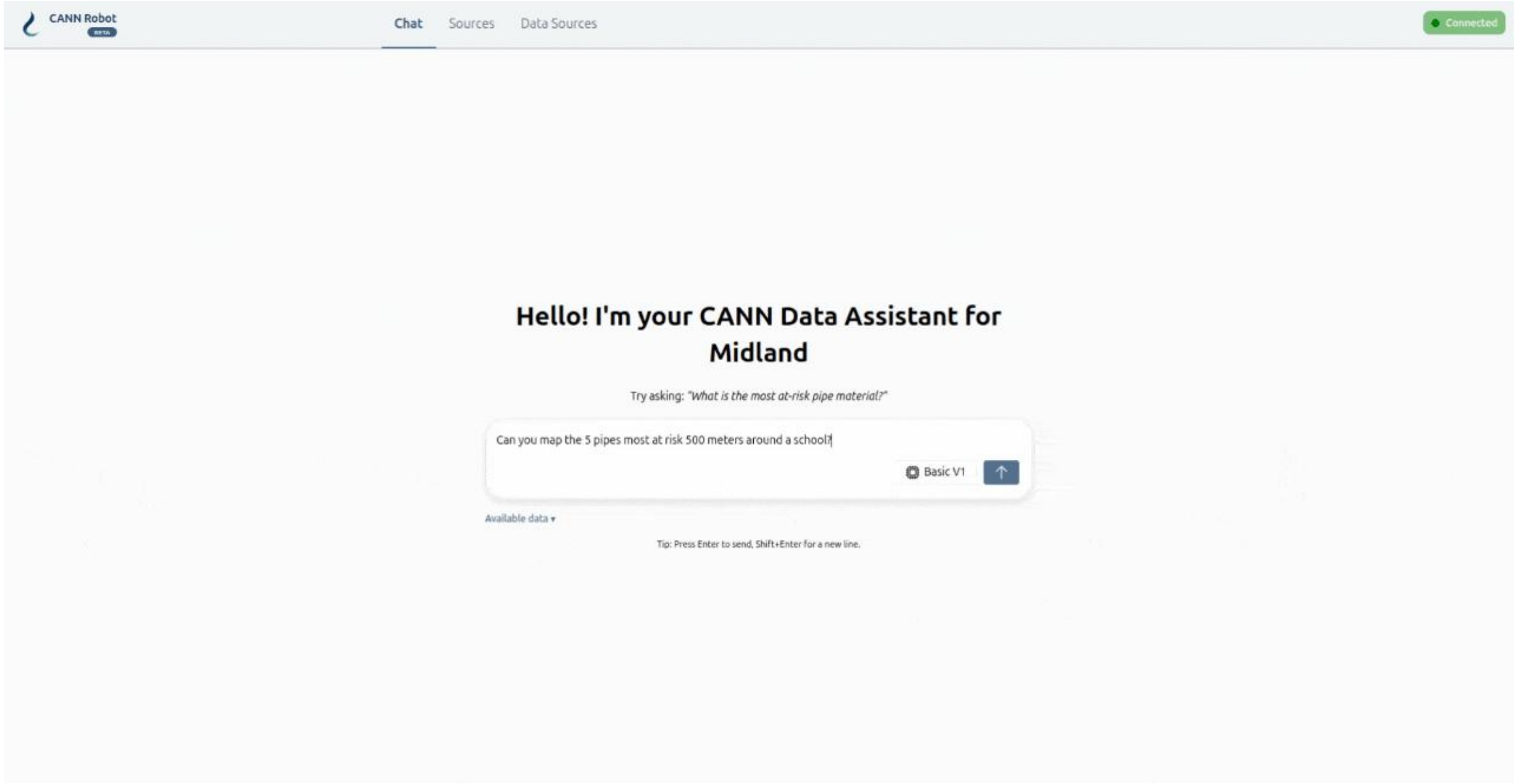
"What is the total length of cast iron pipes installed between 1949 and 1970?"

Unstructured Data Retrieval

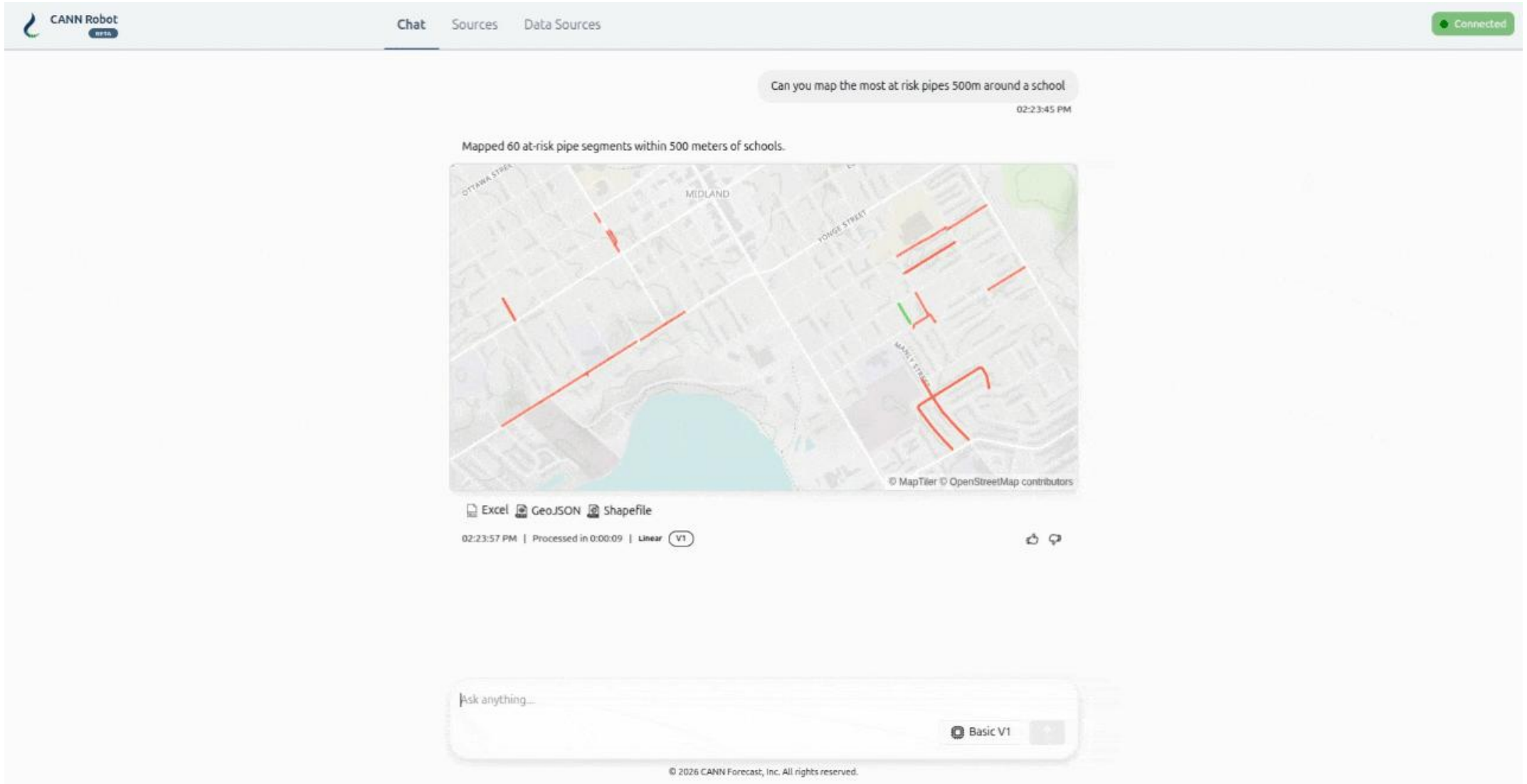
- Process text documents into a knowledge graph
- Store as *documents*, *chunks*, *entities*, and *communities*
- Combine graph traversal with semantic search to retrieve relevant context

"What are the primary methods of disinfection in the water distribution system?"

Chatbot


CANN Robot
BETA
Chat
Sources
Data Sources
Connected
Hello! I'm your CANN Data Assistant for Midland
Try asking: "What is the most at-risk pipe material?"
Can you map the 5 pipes most at risk 500 meters around a school?
Basic V1
Available data
Tip: Press Enter to send, Shift+Enter for a new line.

Chatbot


CANN Robot
BETA
Chat
Sources
Data Sources
Connected
Can you map the most at risk pipes 500m around a school
02:23:45 PM
Mapped 60 at-risk pipe segments within 500 meters of schools.
MIDLAND
YONGE STREET
© MapTiler © OpenStreetMap contributors
Excel
GeoJSON
Shapefile
02:23:57 PM | Processed in 0:00:09 | Linear V1
Ask anything...
Basic V1
© 2026 CANN Forecast, Inc. All rights reserved.

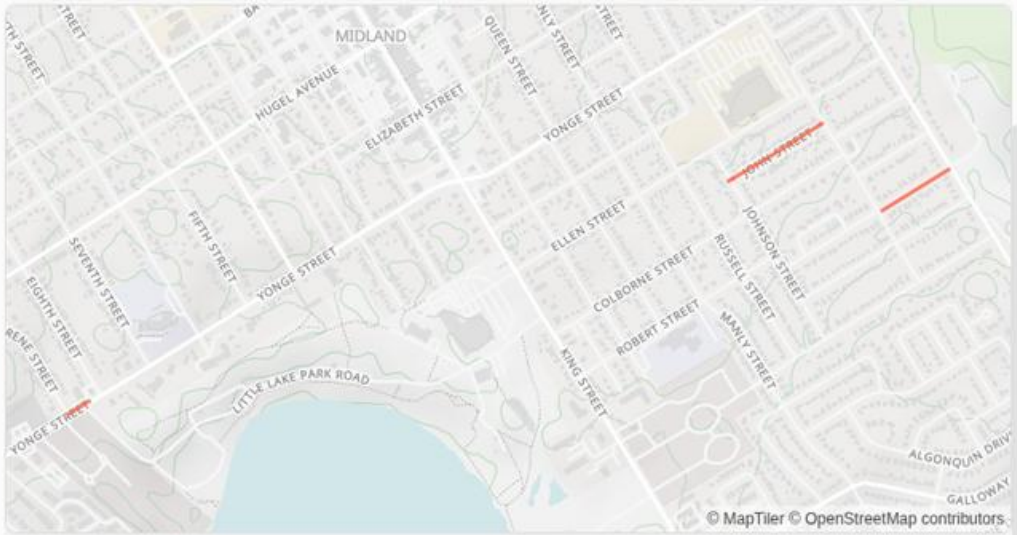
Chatbot

CANN Robot BETA

Chat | Sources | Data Sources

Connected

of failure.



Excel GeoJSON Shapefile

11:31:40 AM | Processed in 0:00:17 | Linear V1

What is the community GHG emission reduction target for the Town of Midland by 2028, relative to 2015 levels?

11:39:29 AM

The community GHG emission reduction target for the Town of Midland by 2028 is a 6% decrease relative to 2015 levels. This equates to an absolute reduction of **8,079 tCO2e**, reducing total community emissions from **134,650 tCO2e (2015)** to **126,571 tCO2e (2028)**.

11:39:37 AM | Processed in 0:00:06 V1 | Most Relevant: *MLCC-Midland-Final-2019-1.pdf (Page 9)* 7 Sources **See all**

Ask anything...

Basic V1

Accuracy Matters — Because Decisions Have Consequences

Tested on 10,000 real-world asset questions

Standard LLM (e.g. ChatGPT	CANN Asset chatbot
Accuracy: **90-92%**	Accuracy: **> 99%**
1 error every 10 answers	Near-zero hallucinations

Reliable AI is essential for operational decision-making

Chatbot

From Data to Decisions

- Utilities already have the data — but it is underused and fragmented
- AI enables continuous, data-driven decision-making
- Moving from:
  - Static plans → Dynamic systems
  - Reactive repairs → Proactive strategies
  - Data → Actionable insights
- **AI is not replacing expertise**

IWA Webinar

Auto-AMP: AI for water Infrastructure Management

CANN FORECAST | Smart Water Management

PRESENTATION 3

Understanding Micropollutant Rejection Mechanism by Polyamide Membranes via Data-Knowledge Co-Driven Machine Learning

Ruobin Dai
Professor
Tongji University, China

Ruobin Dai's research lab focuses on the integration of artificial intelligence with membrane-based water treatment processes.

Supported by national and municipal funding, he has published papers in leading journals, including *Nature Water*, *Sci. Adv.,* and *Environ. Sci. Technol*. He holds 4 U.S. patents; some applied in full-scale water treatment plants.

His work has received multiple awards, including the Gold Medal at the International Exhibition of Inventions Geneva, the 2024 First Prize of the Shanghai Science and Technology Progress Award, the Sustainable Engineering Innovation Award from the Membrane Society of Australasia, Excellent Research Advisor of American Chemical Society, etc.

Beyond research, he serves as an early-career editorial board member for *Desalination*, guest editor for *Sep. Purif. Technol*., and a managing committee member of the IWA Young Water Professionals China Chapter.

The Application of Artificial Intelligence (AI) in Water and Wastewater Treatment

16 April 2026 | IWA Webinar

Understanding Micropollutant Rejection Mechanism by Polyamide Membranes via Data-Knowledge Co-Driven Machine Learning

Ruobin Dai, Hejia Wang, Zhiwei Wang

Tongji University, Shanghai, China

OUTLINE

01 Why data-knowledge co-driven machine learning?

02 Knowledge 1: Feature relationship

03 Knowledge 2: Causal association

04 Knowledge 3: Internet-scale information

05 Take-home messages

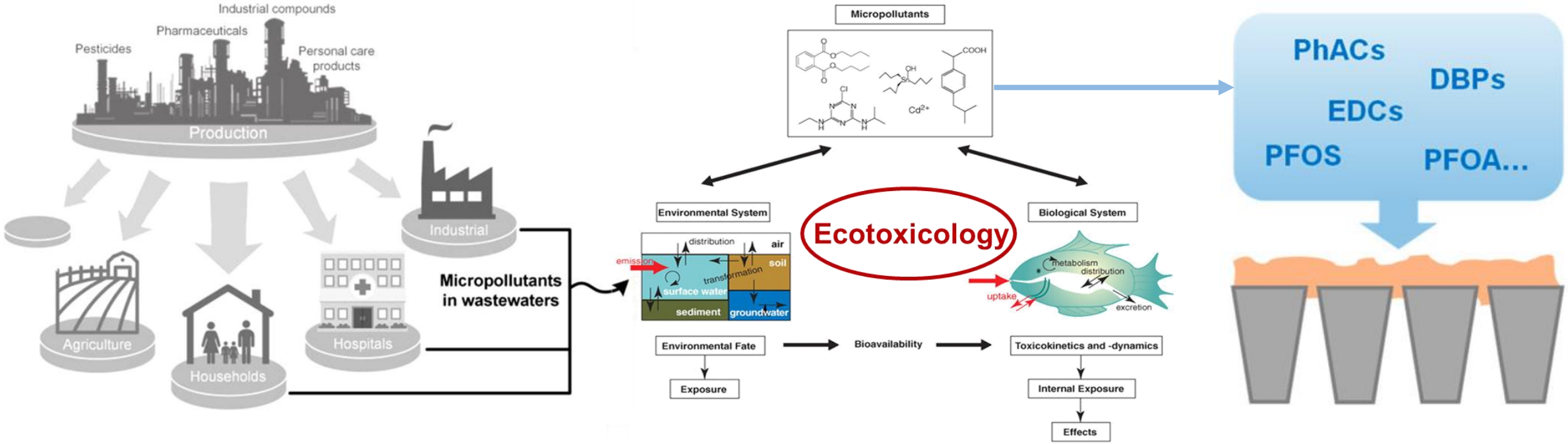
01

PART ONE

Why data-knowledge co-driven machine learning?

Removal of organic micropollutants

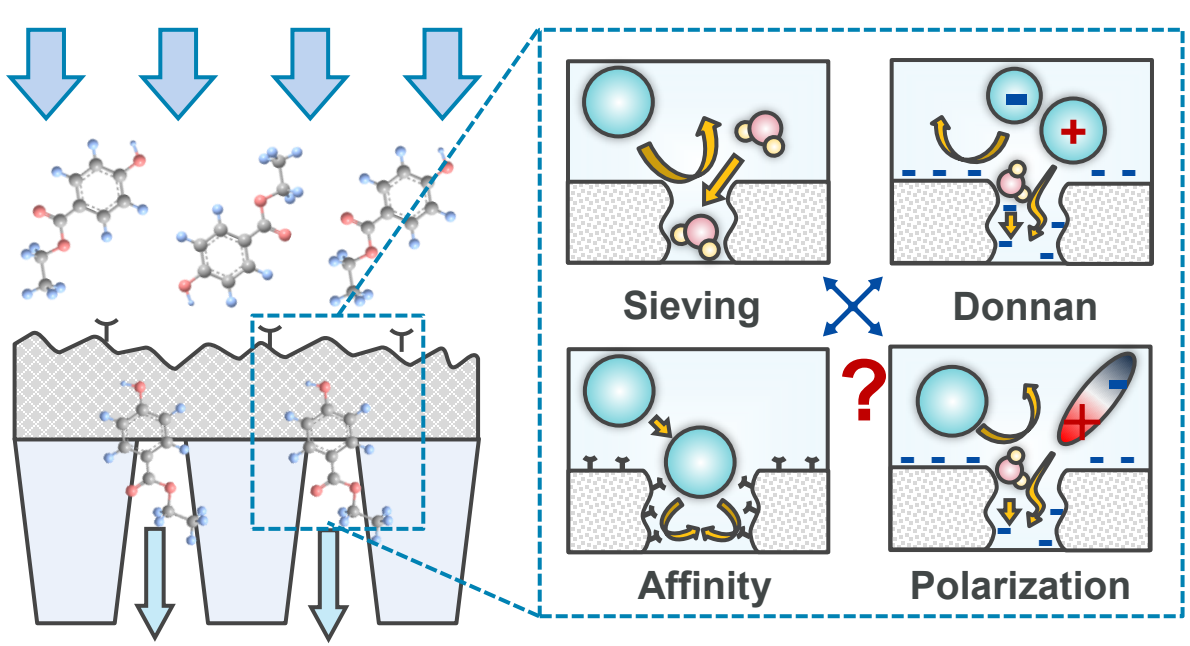
❑ **Efficient removal of organic micropollutants (OMPs) from water has been an urgent issue owing to their potential threats to ecosystem and human health**



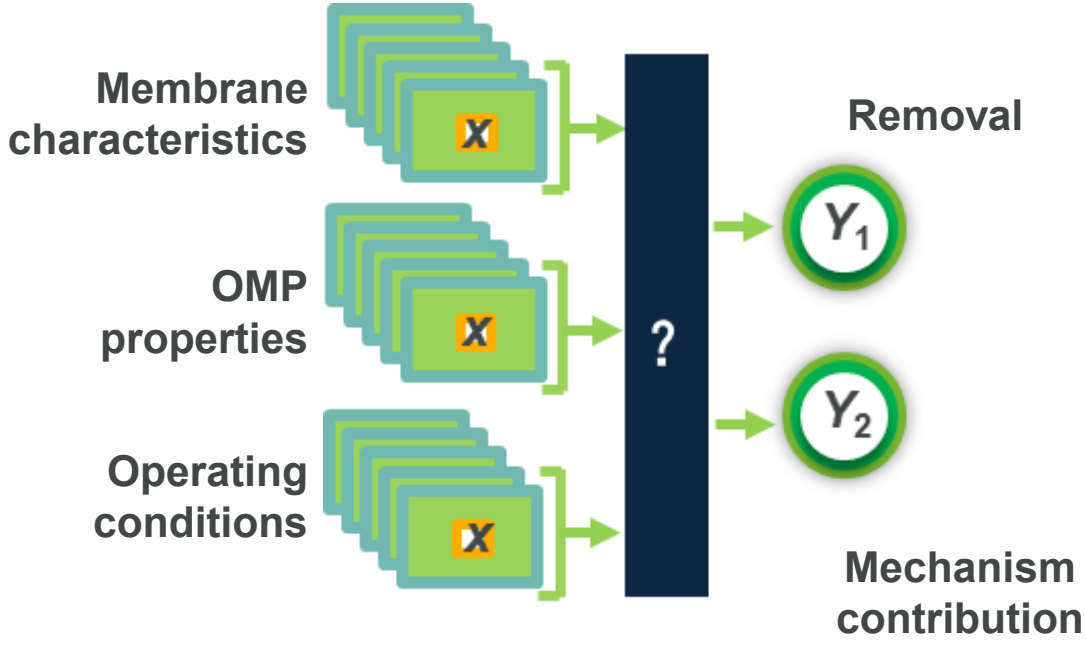
➢ **Polyamide nanofiltration/reverse osmosis membranes have emerged as important candidates for OMP removal due to their exceptional separation efficiency for small molecules**

Complex Membrane Separation Process

- The complexity of membrane separation process as well as various separation mechanisms hinders mechanistic understanding of this system



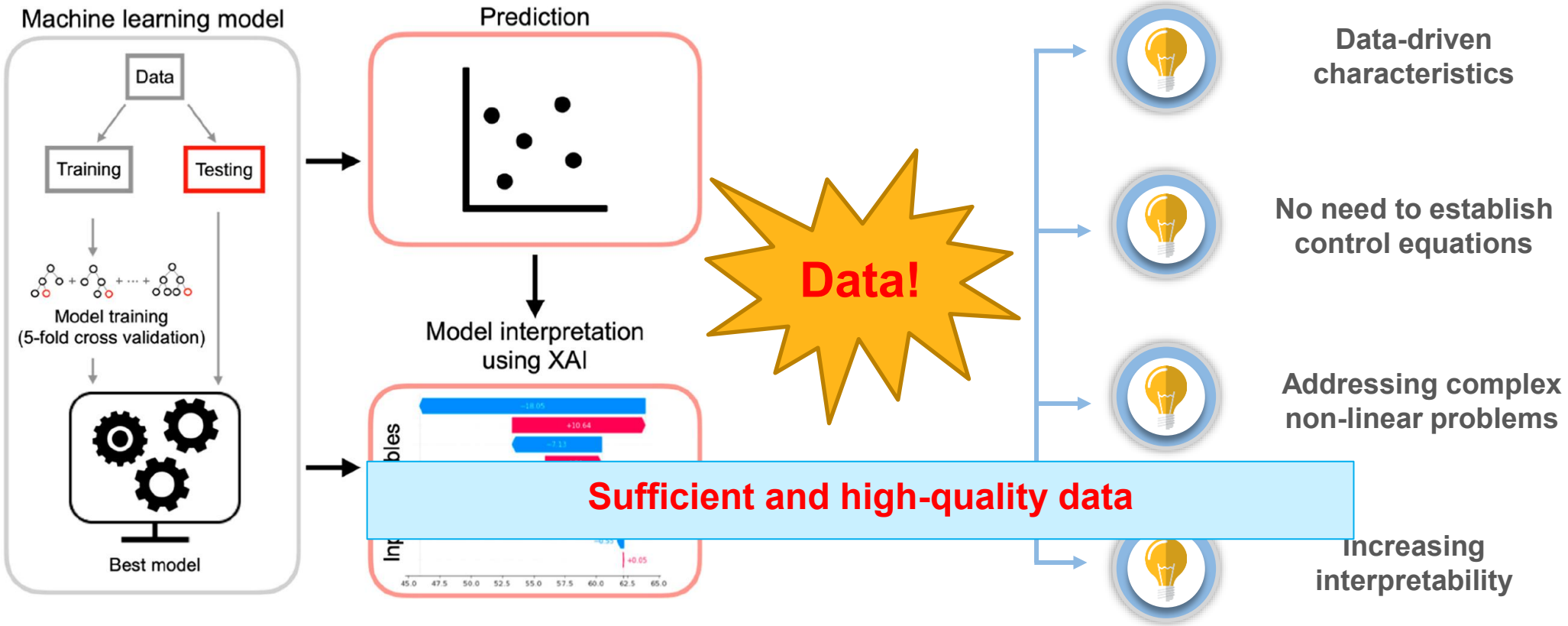
Complex separation mechanisms



Numerous input variables

Machine Learning

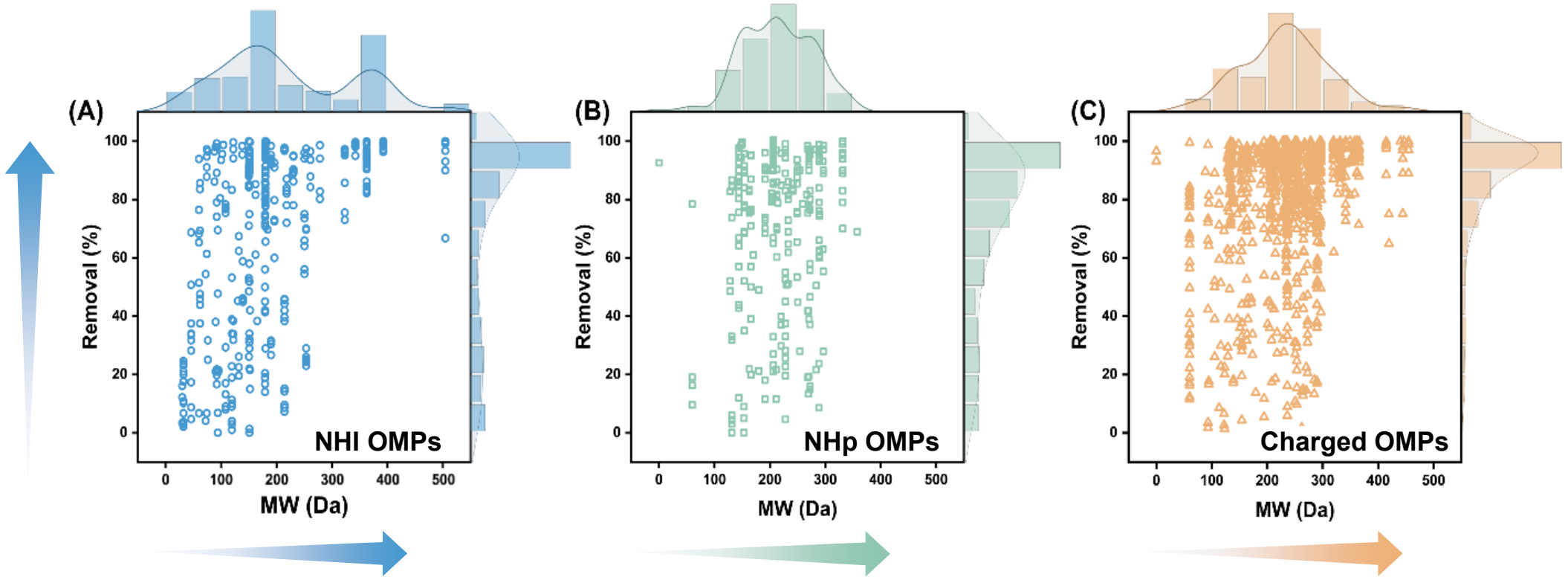
- Machine learning has shown a substantial potential in membrane separation research due to its unique advantages



Environ. Sci. Technol. 2023, 57, 46, 17851–17862; *Environ. Sci. Technol.* 2021, 55, 11348−11359

The challenge of small dataset

- We collected **2438** data points from relevant literature, while this remains a small dataset for modeling the complex membrane separation system



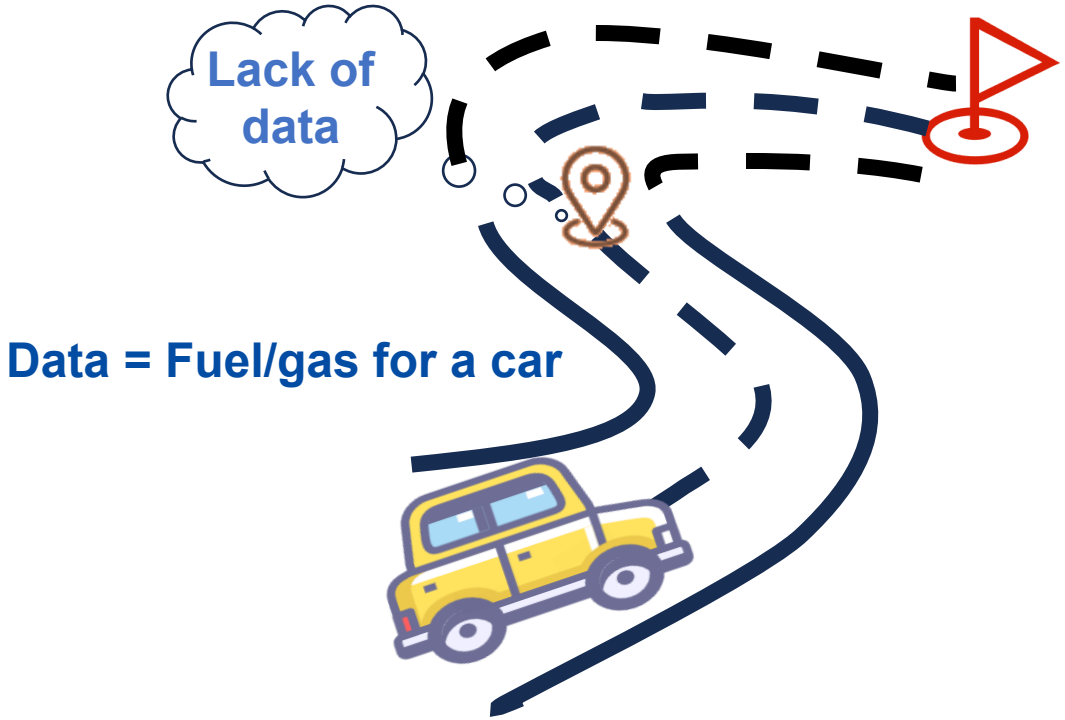
Neutral hydrophilic (NHI) OMPs

Neutral hydrophobic (NHp) OMPs

Charged (Ch) OMPs

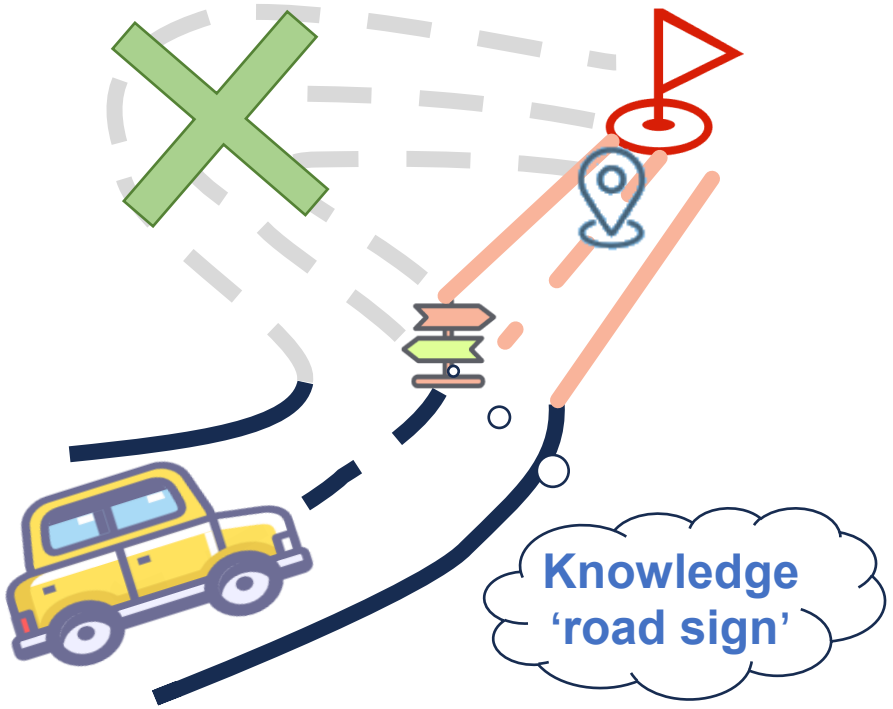
Why Data-Knowledge Co-Driven?

- **Knowledge embedding is an effective approach to bridge the gaps left by data scarcity by integrating domain expertise into the machine learning process**



Data driven

VS



Data-knowledge co-driven

Data-Knowledge Co-Driven Machine Learning

the international water association

Knowledge embedding 1

The embedding of prior feature relationship

Enhance predictive accuracy and understanding of OMP rejection process

Knowledge embedding 2

The embedding of causal associations among features

Investigate the causal pathways and effects on OMP rejection

Knowledge embedding 3

The embedding of internet-scale information

Achieve knowledge transfer of general capabilities to OMP rejection modeling

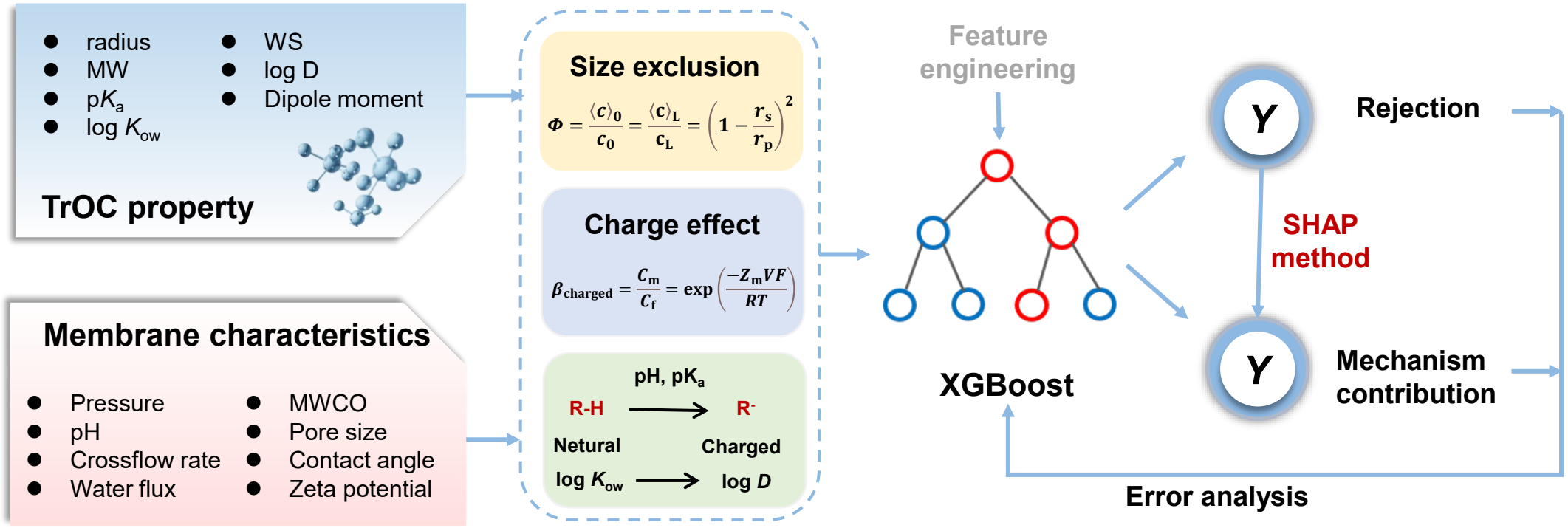
02

PART TWO

Knowledge 1: Feature relationship

2.1 Model Development

- **Quantitatively evaluate rejection mechanisms via Data-knowledge co-driven machine learning**



Dataset construction + **Domain-knowledge embedding** → **Model training** → **Interpretation & deployment**

2.1 Model Development

- Three mathematical models representing different mechanisms were embedded to DKD model by **establishing potential correlations** among basic variables

Size exclusion

$$\Phi = \frac{\langle c \rangle_0}{c_0} = \frac{\langle c \rangle_L}{c_L} = \left(1 - \frac{r_s}{r_p}\right)^2$$

Φ :

- Correlation between **molecular radius** and membrane **pore sizes** was established

Charge effect

Charge product = Zeta × Molecular charge

Charge product :

- Correlation between **molecular charge** and membrane **surface charge** was established

Dissociation

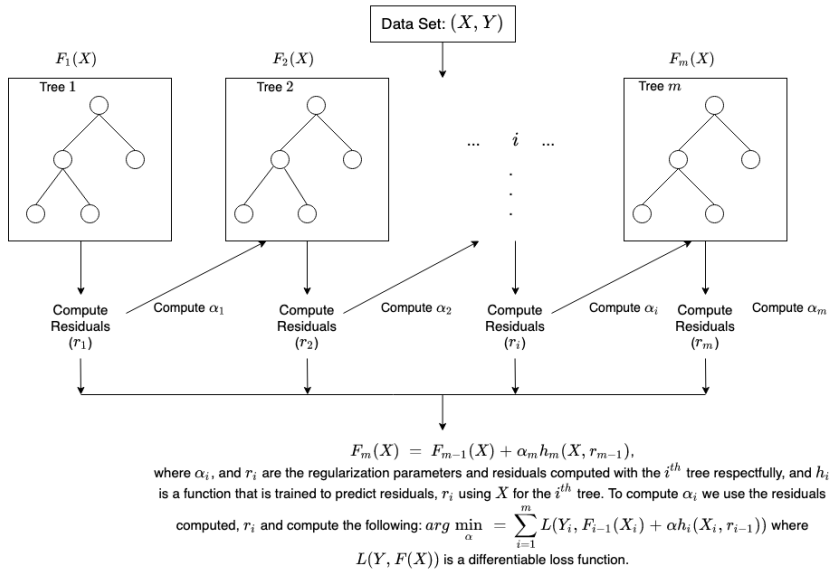
R-H ⟶ R^-

Netural ⟶ Charged

log K_{ow} ⟶ log D

Log D :

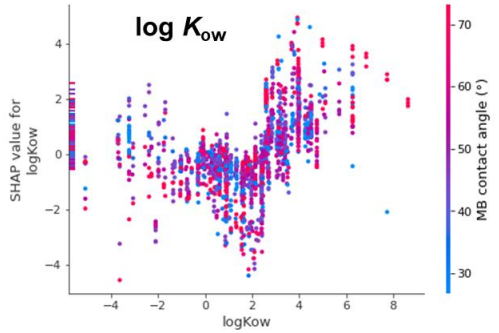
- Correlation between **hydrophobicity** and **molecule dissociation** was established



Conventional models ⟷ **XGBoost ML model**

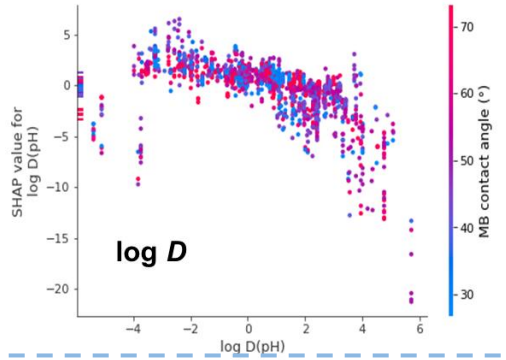
2.2 Model Performance Evaluation

❑ Improved mechanism understanding



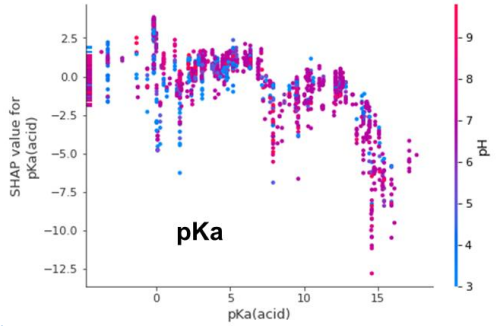
$$\log D = \log\left(\frac{[\mathrm{HA}]_{\mathrm{octanol}} + [\mathrm{A}^-]_{\mathrm{octanol}}}{[\mathrm{HA}]_{\mathrm{water}} + [\mathrm{A}^-]_{\mathrm{water}}}\right)$$

Introduce log *D*: considering alterations in OMP hydrophobicity resulting from the dissociation



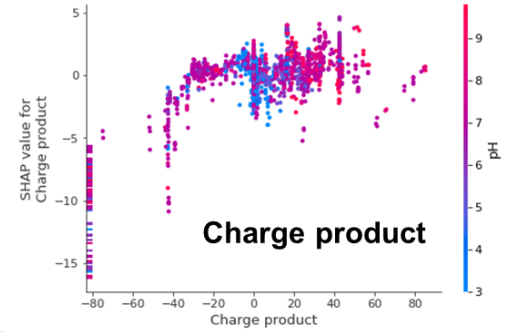
DD model did not identify the effects of hydrophobic interactions correctly

Log *D* (hydrophobicity) exhibited a negative correlation with the rejection



charge product = molecular charge × zeta potential

Introduce Charge product: combining the impact of molecular charge and membrane potential

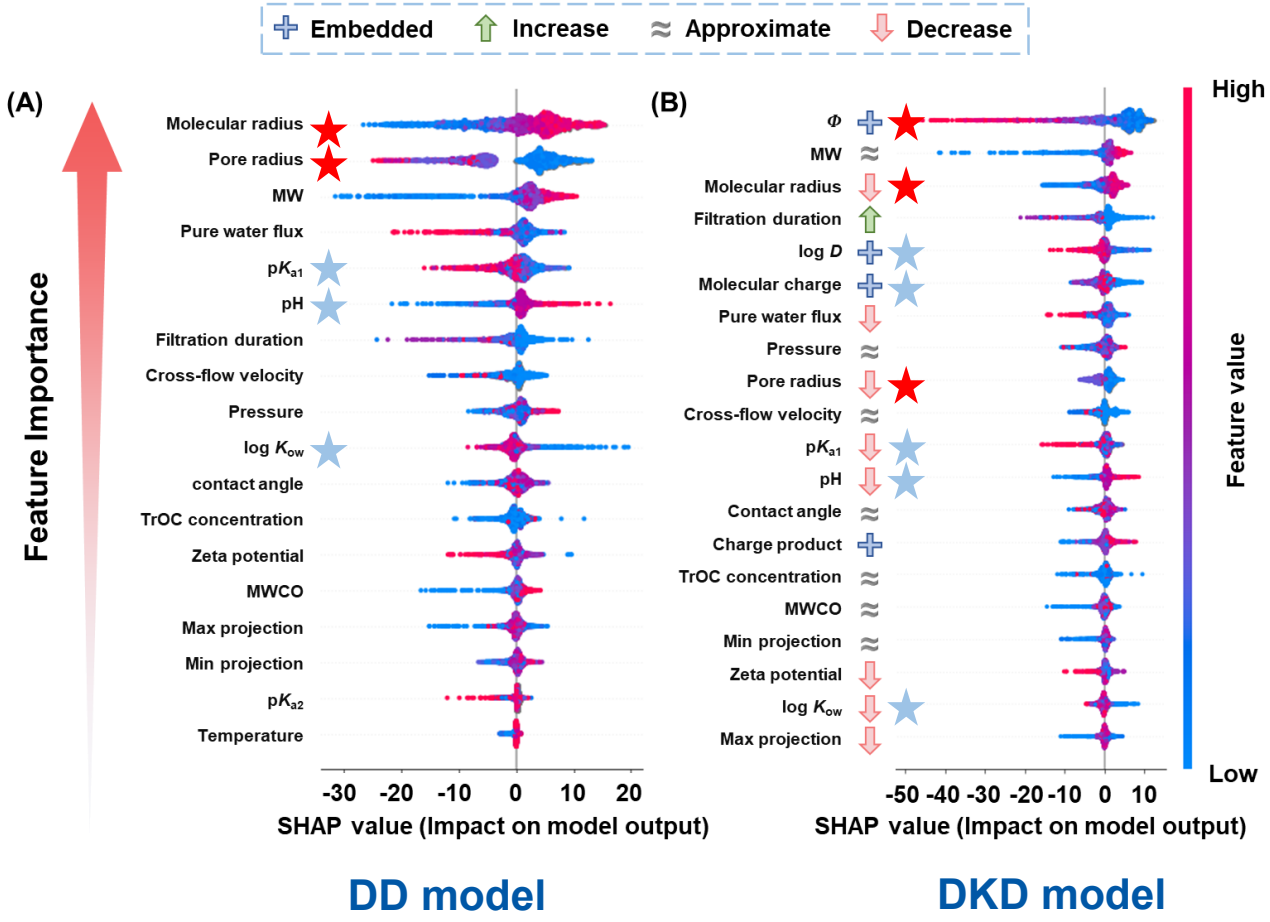


DD model did not reveal the effect of molecular charge and zeta potential

When the Charge product is negative, MP rejection decreased

2.3 Model Interpretation

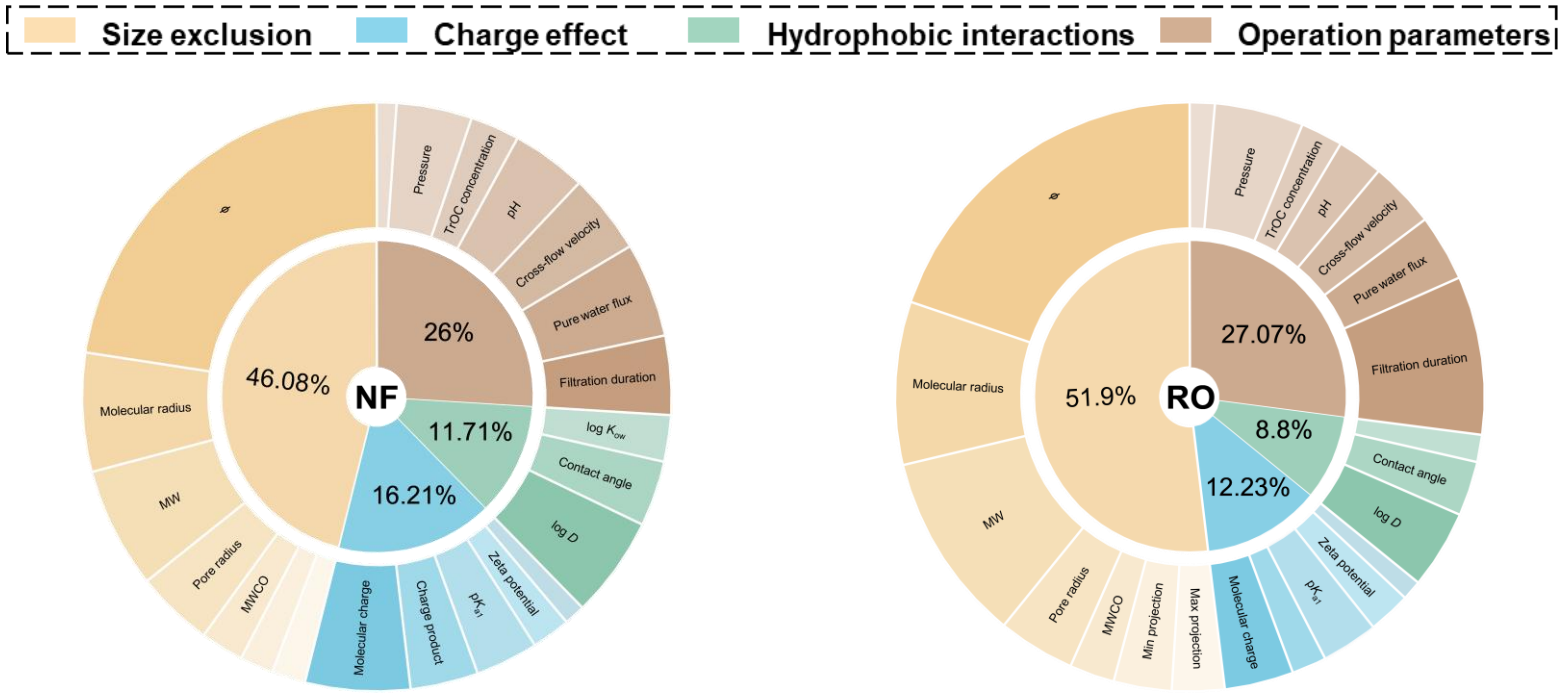
❑ SHAP summary plot of the two models



- Notable change in feature importance ranking was observed after domain-knowledge embedding
- DKD model decision relies more on Φ rather than molecular radius and Pore sizes
- The rankings of pH, pK_a, and log K_{ow} decreased, with log D eclipsing their importance

2.4 Mechanism contribution

❑ Comparison of rejection mechanisms between NF and RO membranes



- ➢ **Size exclusion** plays a dominant role in the rejection of OMPs by NF and RO membranes
- ➢ Compared with RO, **NF relies more on charge effect and hydrophobic interactions**

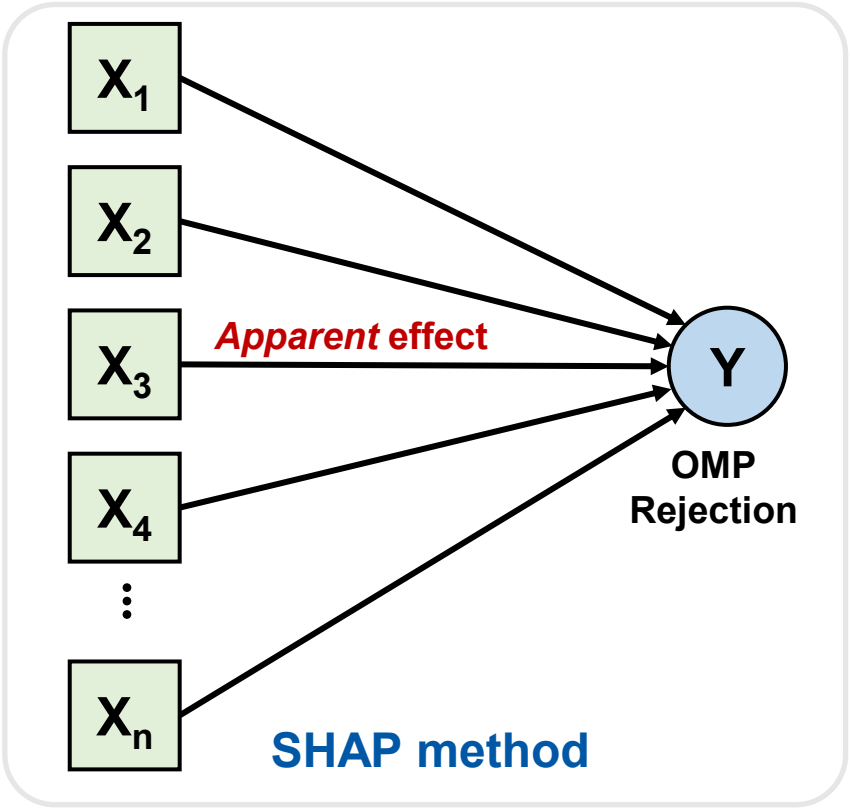
03

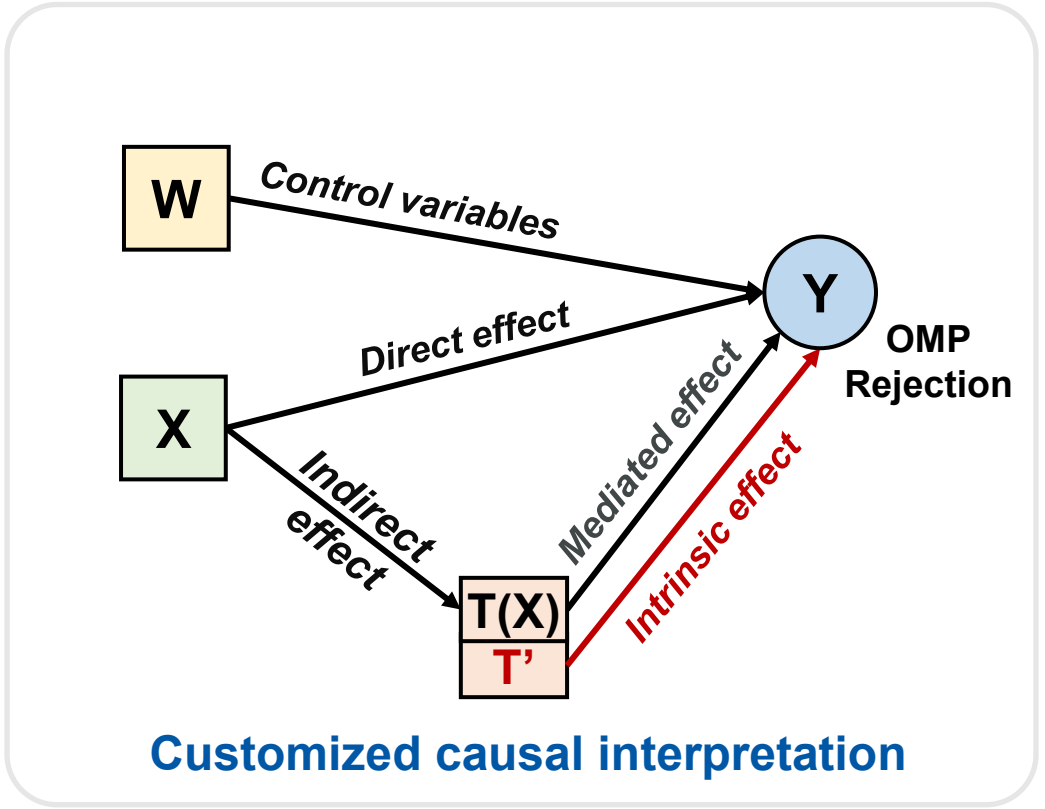
PART THREE

Knowledge 2: Causal associations

3.1 Causal Machine Learning

❑ Beyond correlation, how to answer "If-then" questions?

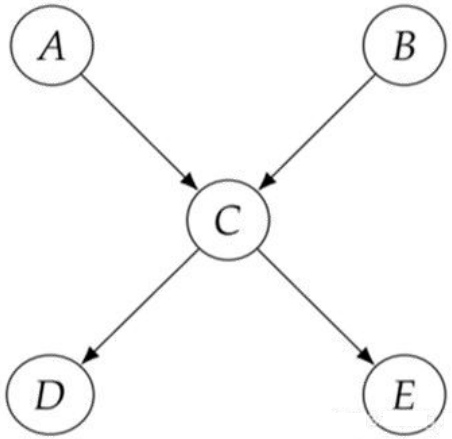


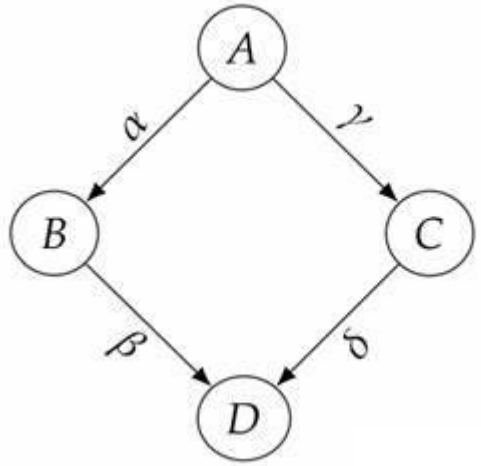


- Correlation-based methods only reveal the apparent effect of features
- Customized causal interpretation can distinguish the indirect and direct effects

3.1 Causal Machine Learning

- Causal machine learning has advanced to the stage of causal relationship evaluation, based on the prior causal relationship acquired from experimental findings





Causal relationship discovery

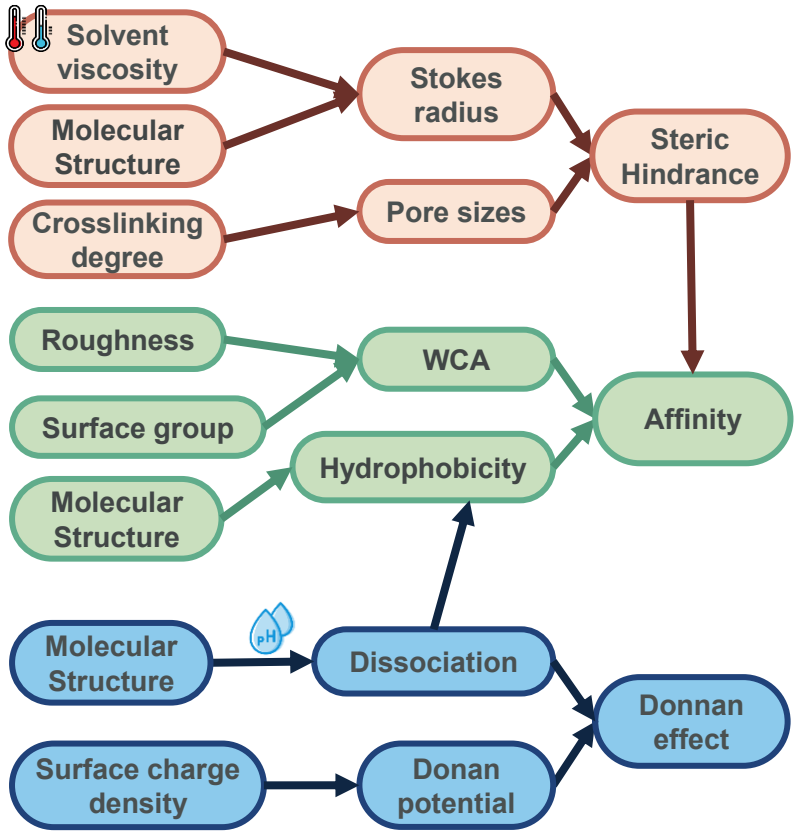
- Target: Unravel the causal relationships between variables
- Do causal relationships exist between pairs of variables?

Causal effect inference

- Target: Quantify the causal effect between variables
- How much will the sales decrease if the price increases by 1 dollar?

3.1 Causal Machine Learning

- Causal inference framework can be customized based on prior variable associations to decouple the effect different rejection mechanism



Prior causal relationship (Knowledge)

$$Y = \theta(X) \cdot T + g(X, W) + \varepsilon$$

$$\mathbb{E}[\varepsilon | X, W] = 0$$

$$T = f(X, W) + \eta \qquad \mathbb{E}[\eta | X, W] = 0$$

$$Y - \mathbb{E}[Y|X, W] = \theta(X) \cdot (T - \mathbb{E}[T|X, W]) + \varepsilon$$

$$q(X, W) = \mathbb{E}[Y|X, W]$$

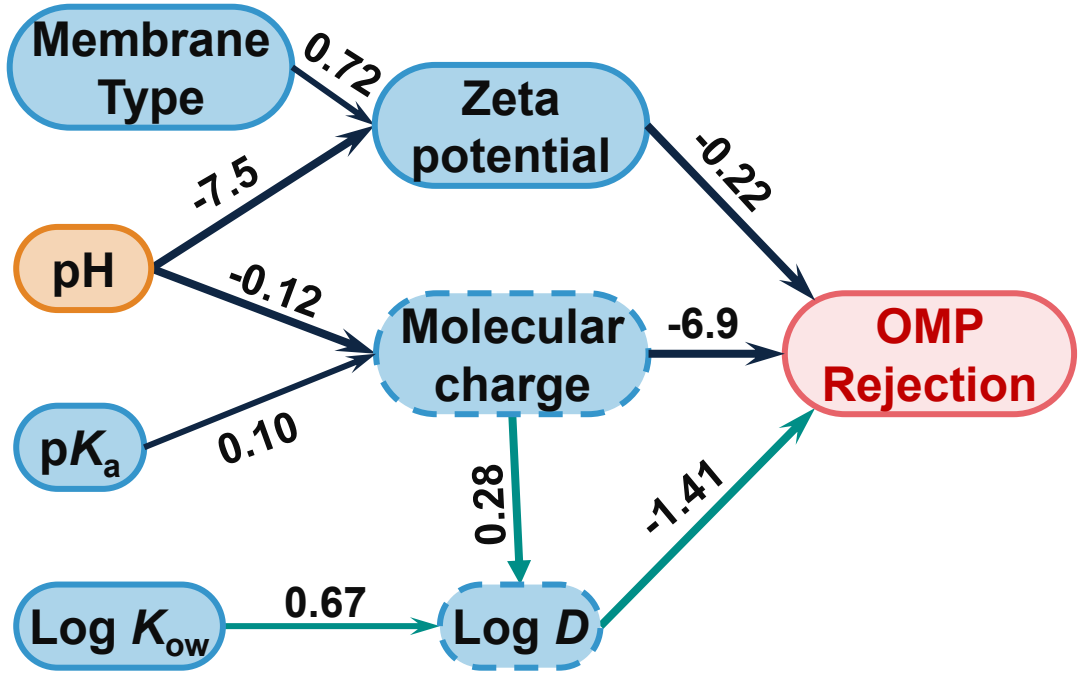
$$f(X, W) = \mathbb{E}[T|X, W]$$

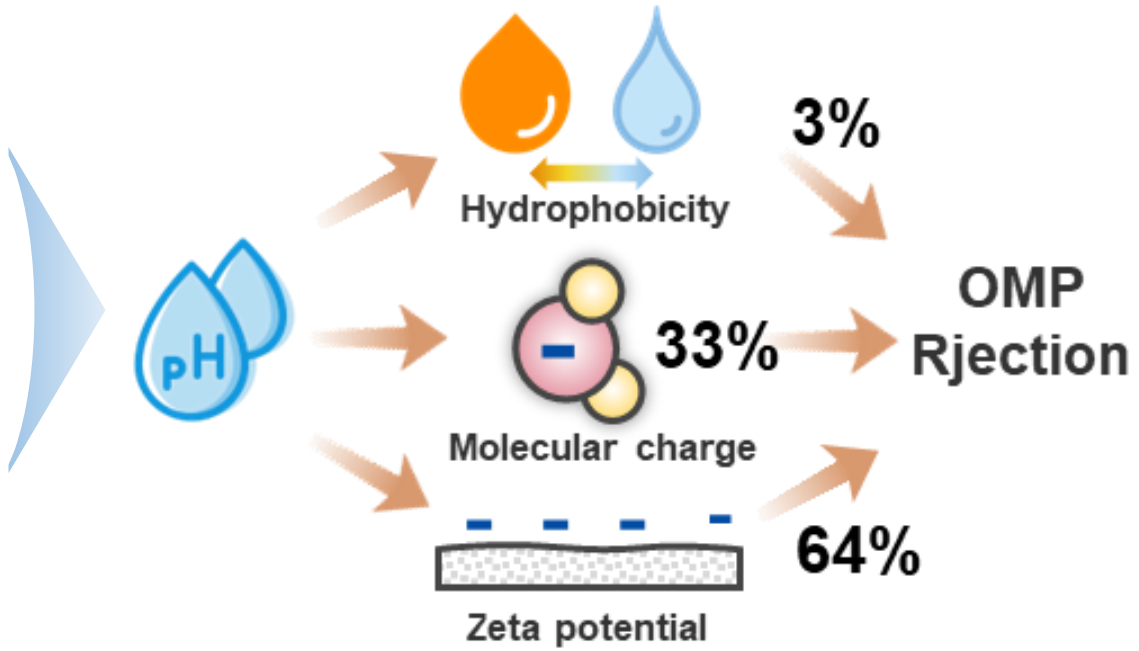
$$\tilde{Y} = \theta(X) \cdot \tilde{T} + \varepsilon$$

Customized causal inference framework

3.2 Causal Pathways and Effects

❑ The mechanisms and contributions of pH in influencing OMP rejection

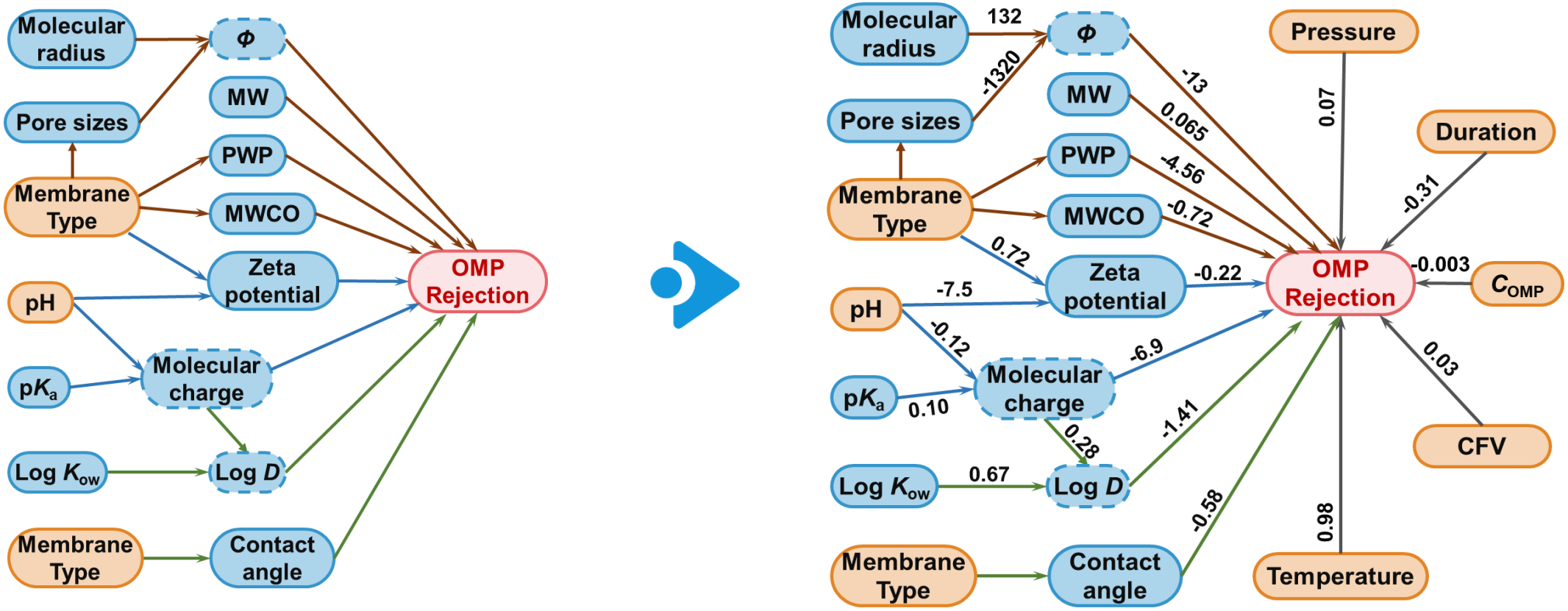




- pH influences the OMP rejection by altering molecular charge, hydrophobicity, and zeta potential, contributing 33%, 3%, and 64%

3.3 Membrane System Optimization

❑ Causal graph of system variables affecting OMP rejection



- The causal graph linking the variables to OMP rejection was constructed through customized causal interpretation

04

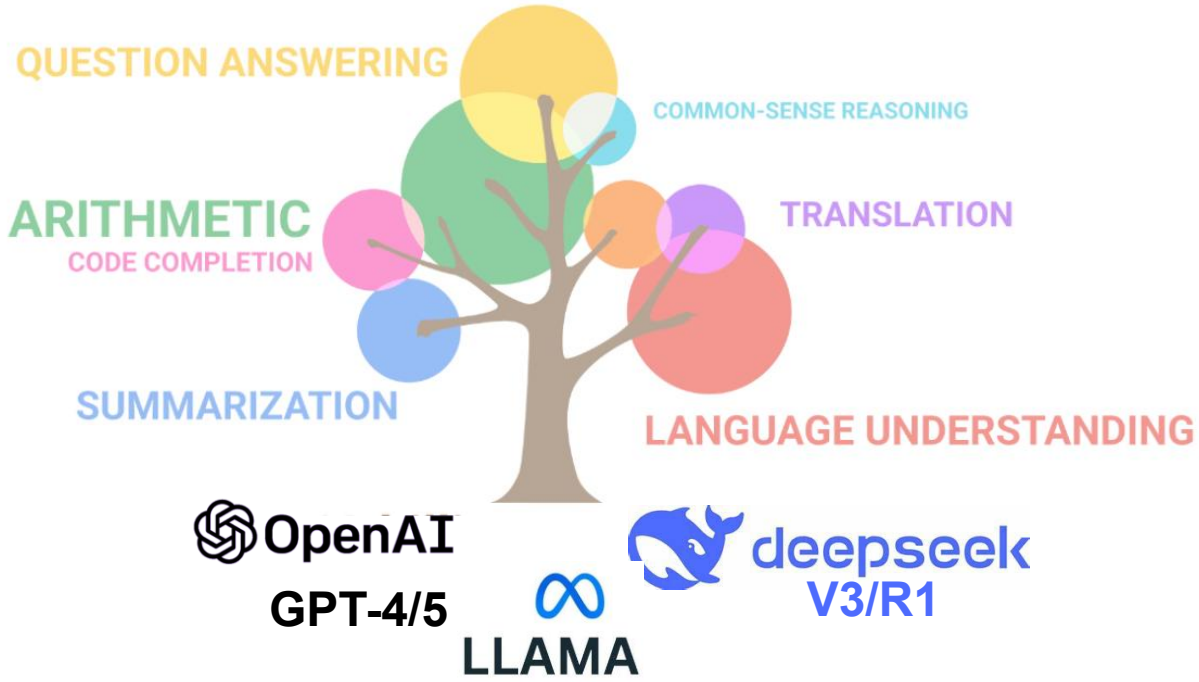
PART FOUR

Knowledge 3: Internet-scale information

4.1 Large Language Models

- LLMs, trained with internet-scale information, has demonstrated cross-domain application capabilities due to the emergent intelligence
- These knowledge can be transferred to environmental tasks through fine-tuning

Model	Parameter scale	Dataset scale
GPT-1	0.12 B	5 GB
GPT-2	1.5 B	40 GB
GPT-3	175 B	45 TB
GPT-3.5	175 B	100+ TB
GPT-4	~200 B	100+ TB

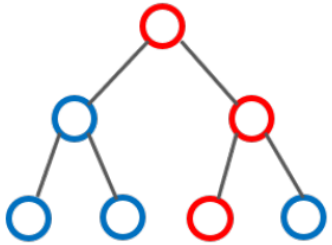


Are LLMs the endgame for membrane process?

Membrane separation process

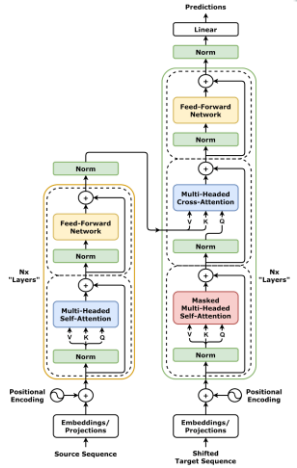
Handgun

Conventional ML Models

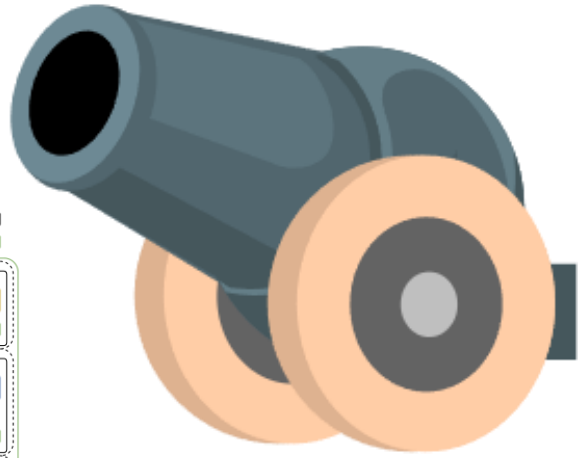


OMP rejection prediction

VS



Cannon

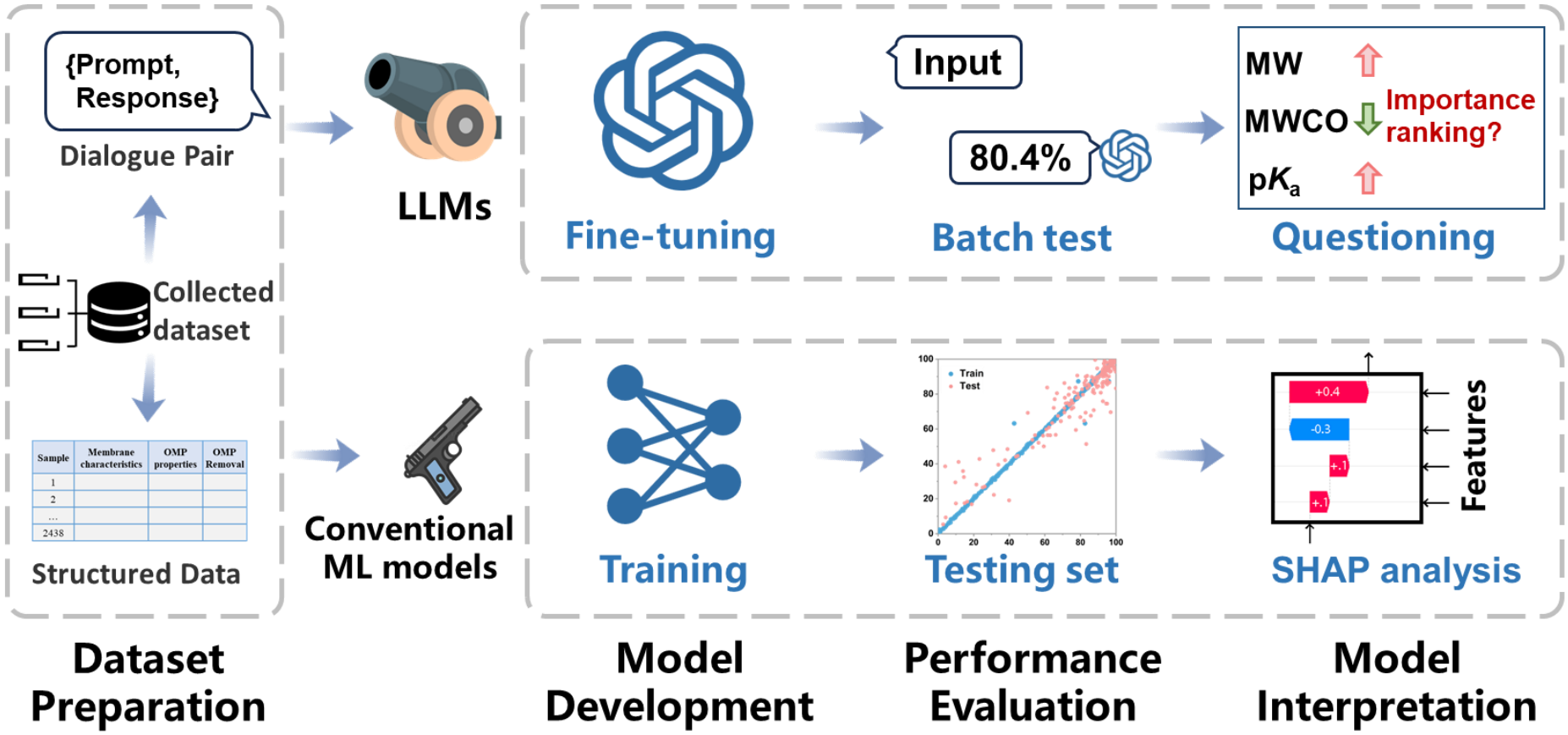


Large Language Models

Which can hit the bullseyes？

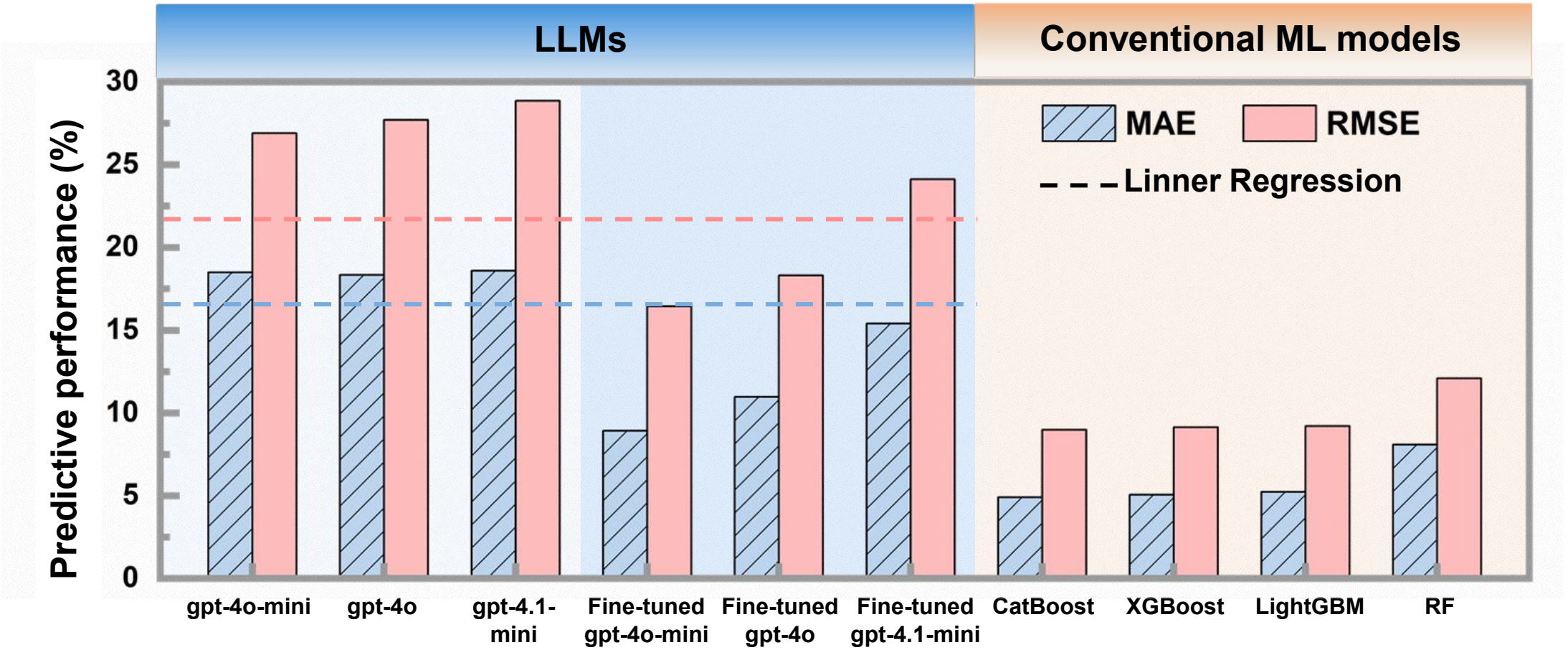
4.2 Workflow of This Study

- We developed a process for fine-tuning and interpreting LLMs, using conventional ML models as a control group



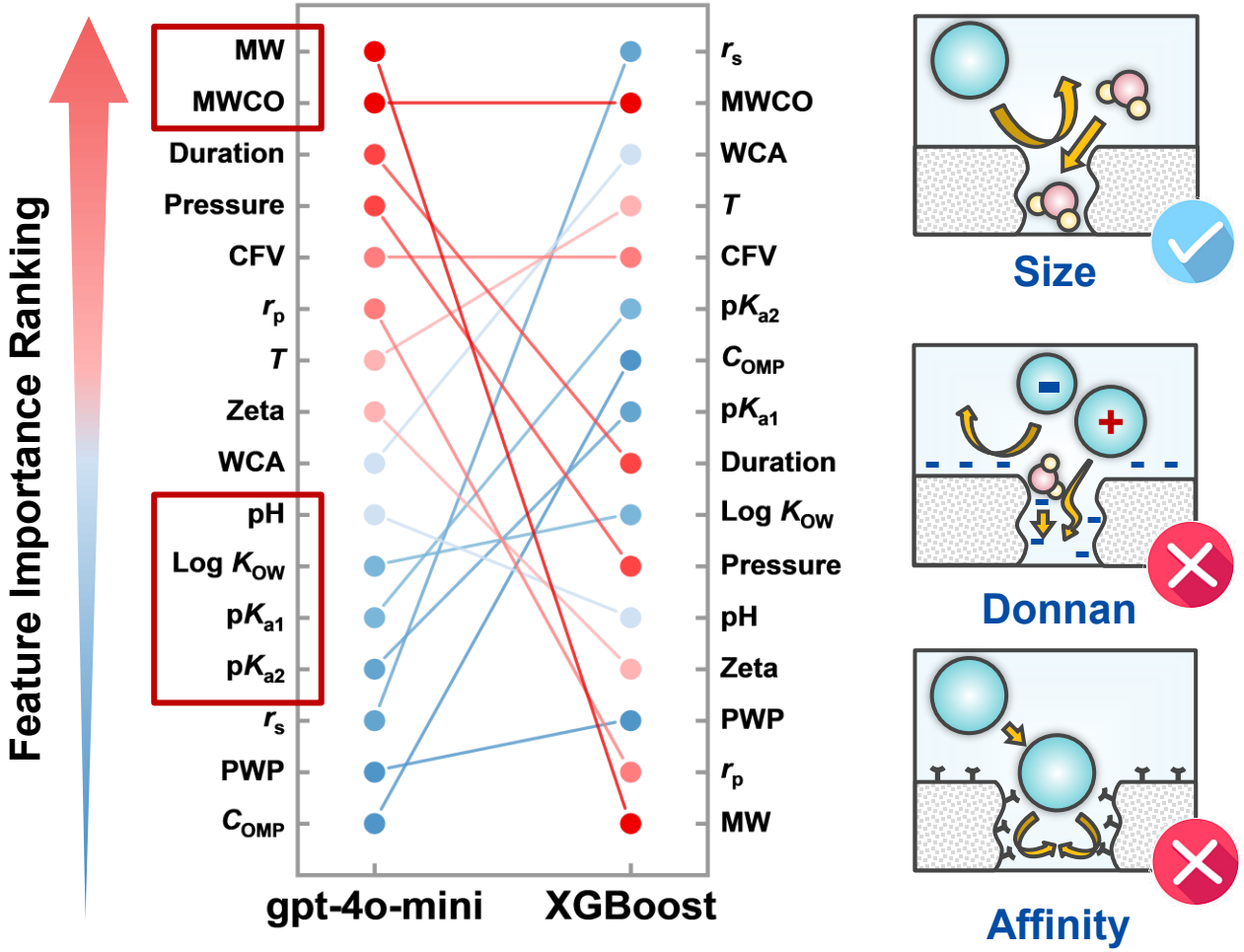
4.3 Predictive Accuracy of LLMs

- The performance of LLMs substantially improved after fine-tuning, while they still slightly underperformed conventional ML models



4.4 Mechanistic Understanding of LLMs

- Despite enhanced comprehension of size exclusion, fine-tuned LLMs sill lack adequate understanding of Donnan effect and solute–membrane interactions



- The Transformer architecture is more effective at capturing sequential dependencies in text than at handling structured numerical features
- Converting structured data into text sequences leads to information loss and redundancy

Larger model ≠ Better model

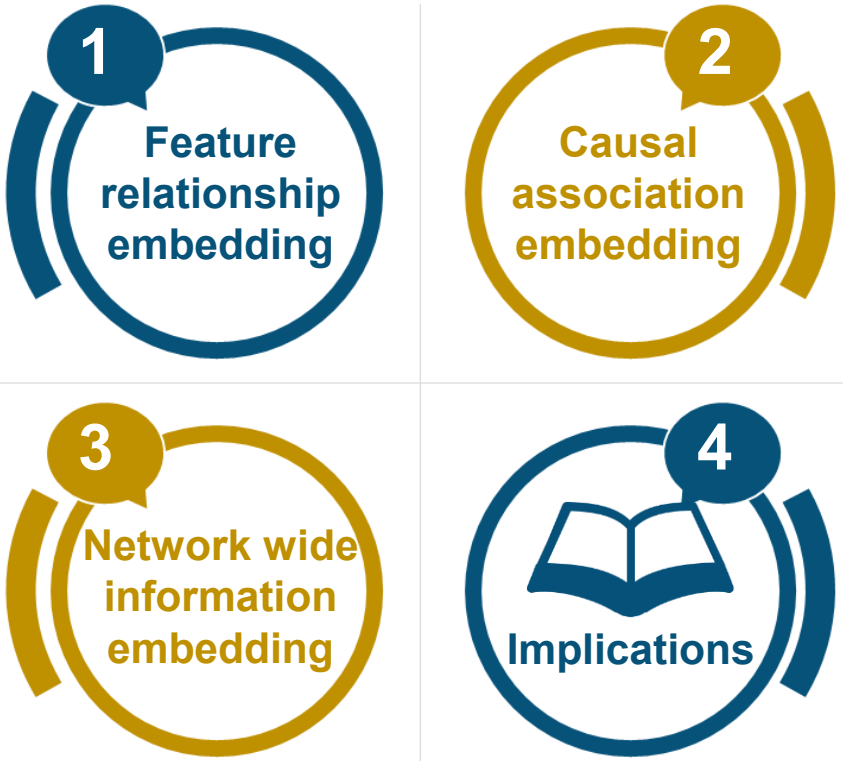
05

PART FIVE

Take-home messages

Take-home Messages

- The contribution of rejection mechanisms in diverse membranes and OMPs were quantitatively evaluated through domain-knowledge embedding
- We demonstrates the potential of LLMs for regression-based modeling and highlights feasible strategies for further improvement



- Based on prior causal associations, the mechanism pathways and causal effects influencing OMP rejection were analyzed
- Data-knowledge co-driven machine learning approaches have broad potential applications in membrane separation modeling

Acknowledgement

- National Natural Science Foundation of China
- Shanghai Municipal Science and Technology Commission

Thanks for your attention!

Q&A

Ruobin Dai

dairuobin@tongji.edu.cn

UPCOMING IWA WEBINARS & EVENTS

Learn more at
https://www.iwa-network.org/learn/pfas-in-the-industrial-water-cycle-a-pragmatic-approach-on-legislation-prevention-and-treatment-1

UPCOMING IWA WEBINARS & EVENTS

Learn more at
https://www.iwa-network.org/learn/metagenomics-and-metabarcoding-to-understand-microbial-dynamics-in-wastewater-treatment-systems

UPCOMING IWA WEBINARS & EVENTS

Learn more at
https://iwa-let.org/

UPCOMING IWA WEBINARS & EVENTS

Learn more at
https://worldwatercongress.org/

UPCOMING IWA WEBINARS & EVENTS

Learn more at
https://digitalwatersummit.org/

JOIN OUR NETWORK OF WATER PROFESSIONALS!

the international water association

IWA brings professionals from many disciplines together to accelerate the science, innovation and practice that can make a difference in addressing water challenges.

SCAN THE QR CODE

Use code **WBNR26IWA20**

for a **20% discount off** new membership.

Join before 31 December 2026

inspiring change

the international water association

Learn more at

http://www.iwa-network.org/iwa-learn/