Cloud-based Open-source Digital Twin for BioChemical Pilot Plant

Trung Trinh, PhD
Research Scientist , SINTEF Industry, Norway
Roman Tschentscher, PhD
Senior Research Scientist , SINTEF Industry, Norway

Contact: trung.trinh@sintef.no

Advances in the digitalisation of the process industries conference
Manchester, Oct 2025

Teknologi for et bedre samfunn

Contents

- Introduction to SINTEF
- Digital Twin Concept for BioChemical Pilot Plant
- Cloud Architecture
- Machine Learning Models
- Big Data Analytics & Visualization
- Conclusion

SINTEF
75 år

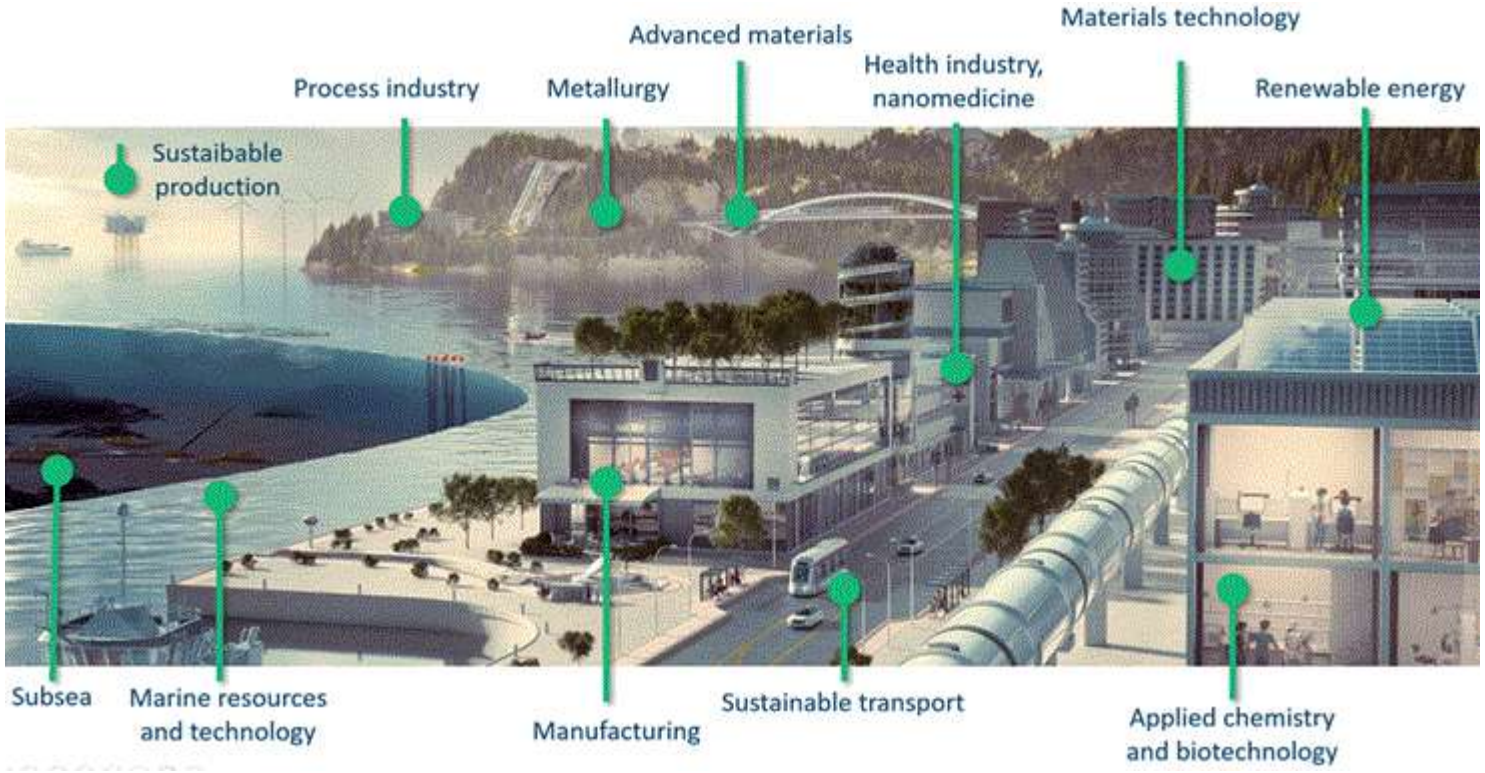
SINTEF

Norwegian Research Organization founded in 1950

INSTITUTTER

- SINTEF Community
- **SINTEF Industry**
- SINTEF Digital
- SINTEF Ocean
- SINTEF Energy

AI in Domain



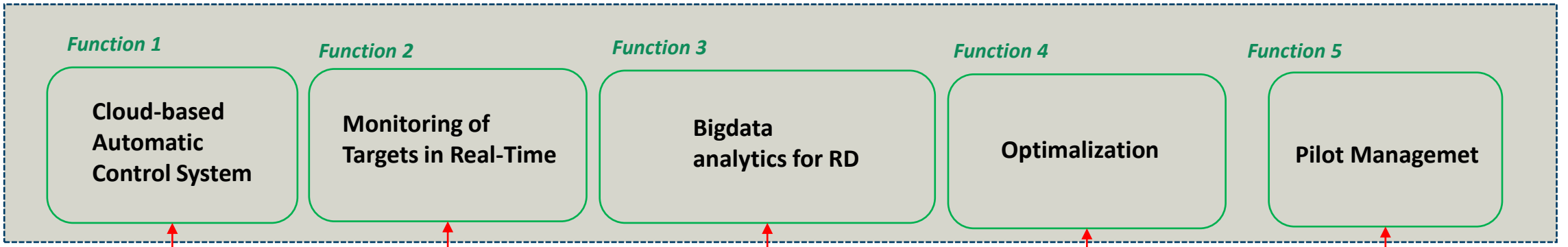
SINTEF

- 2200 employees
- 80 nationalities
- 3300 customers

Digital Twin For Process Pilot

The Digital Twin:
AI, big data, autonomous systems, machine learning, *and* ***cloud solutions***

Data Platform on Azure VM



Pilot /Process

Sytem Control

Data Storage

Process Unit

ML Models

The Digital Twin components:

- ***Pilot***
- ***Data Platform***
- ***ML Models***

Tools for Digital Twins

Kafka app: realtime stream data from Raman analysis equipment to data platform

MQTT app: real-time communication between Cloud and control system of pilot

PostgreSQL: TimescaleDB to collect realtime data from Raman and operation

Python: Data processing linked to ML models

Machine learning models (PLSr, XGBoost, Baysiean Optimalization)

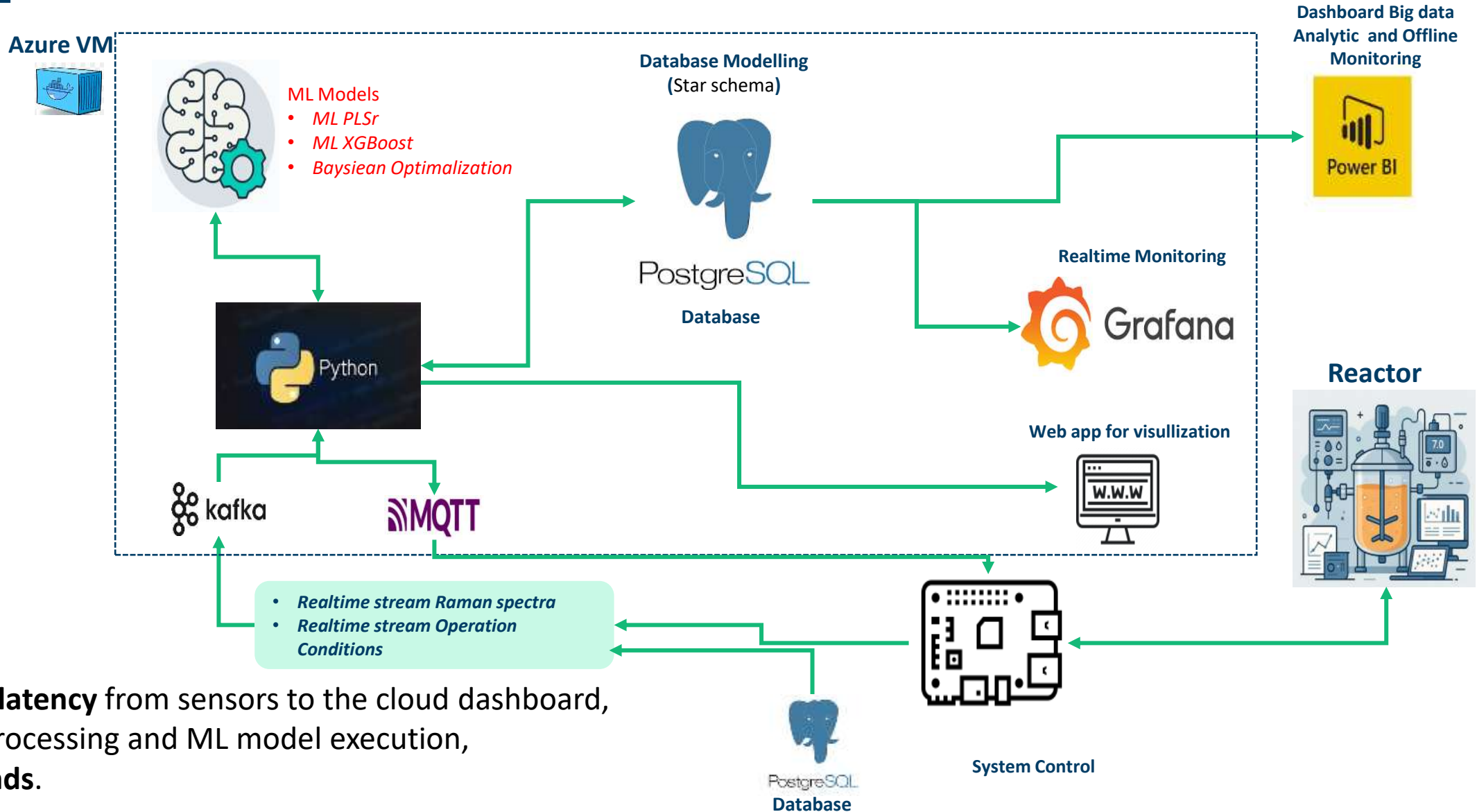
Flask app: web app for visullization

Grafana app: Real time monitoring

Power BI app (Microsoft- license app): big data analytics

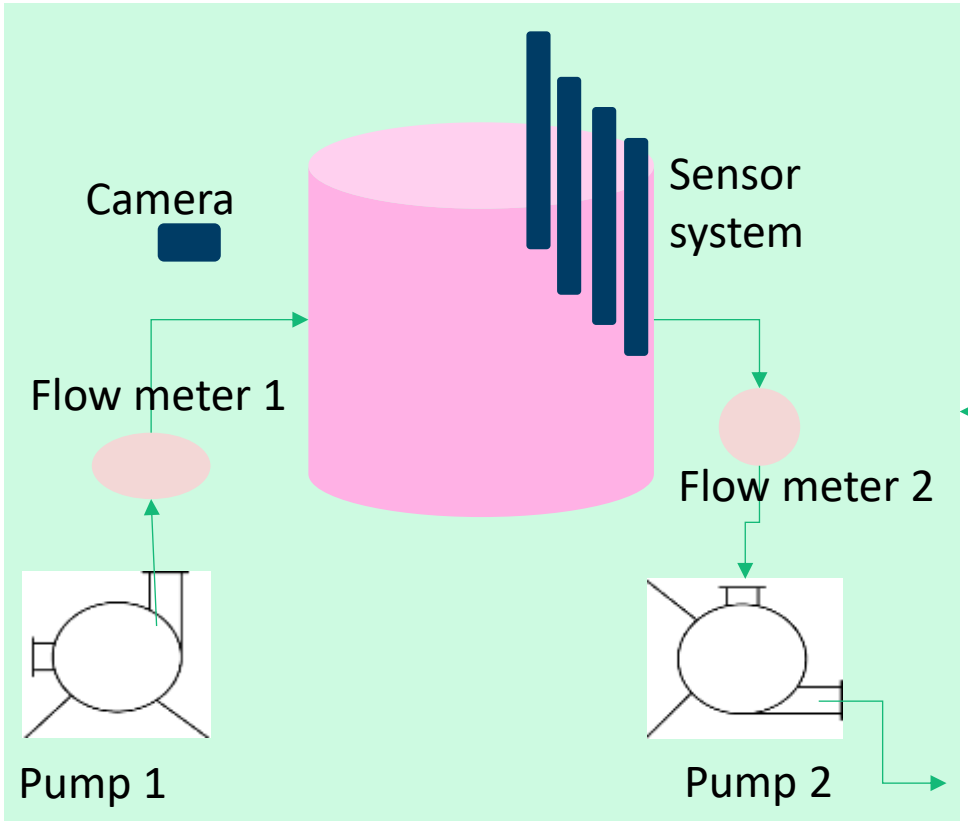
Docker app: an environment for installing, running, and packaging all applications within a data platform.

Cloud Architecture for The Digital Twins

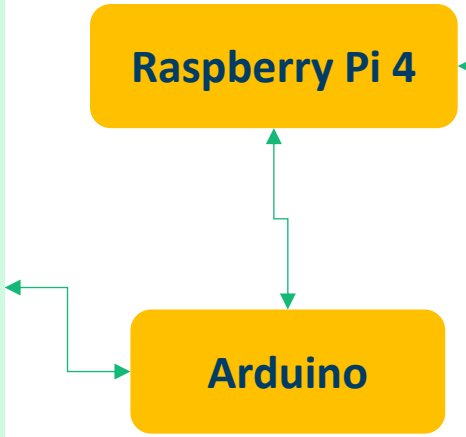


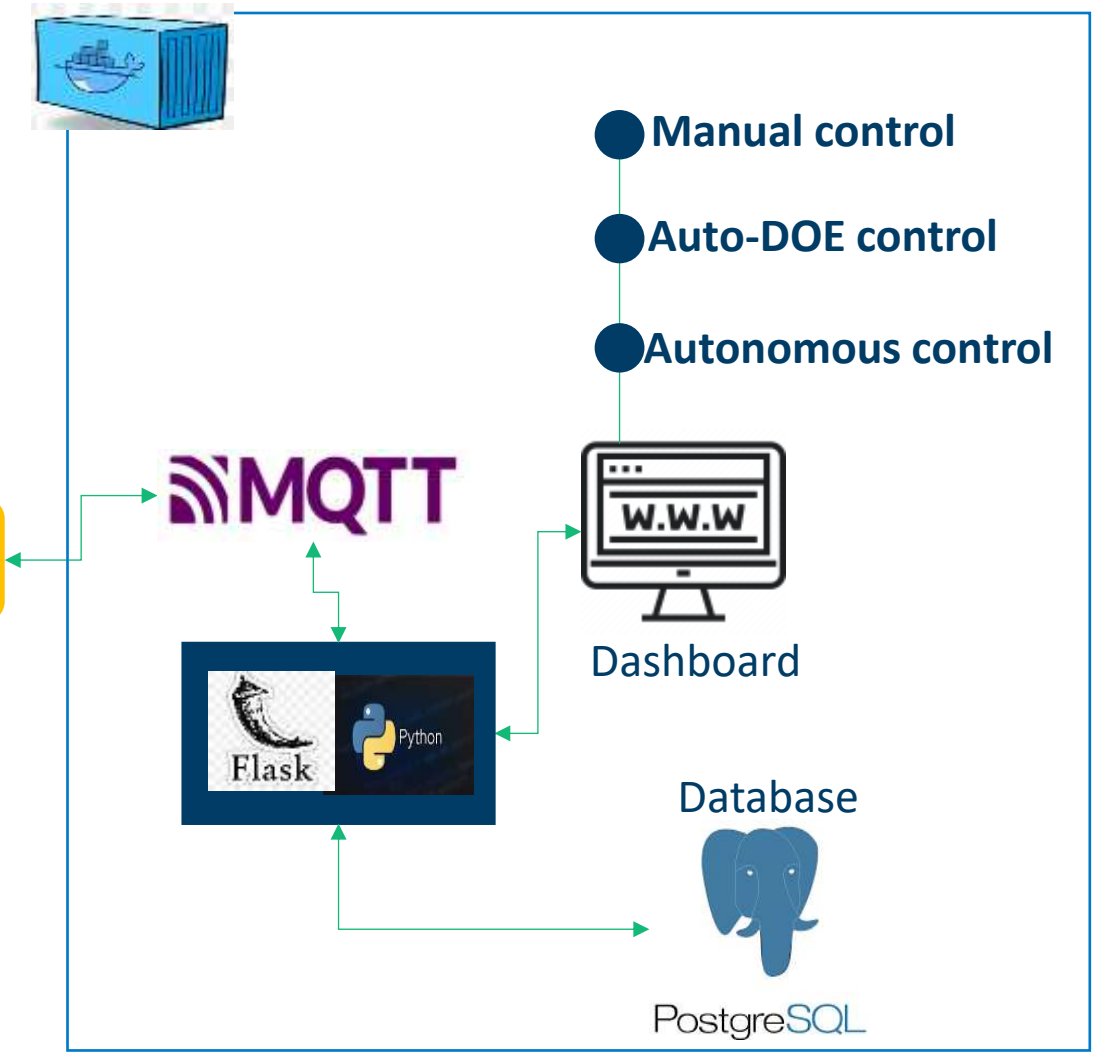
The **end-to-end latency** from sensors to the cloud dashboard, including data processing and ML model execution, is **below 2 seconds**.

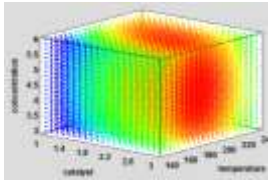
Cloud based Pilot System Control



Reactor System





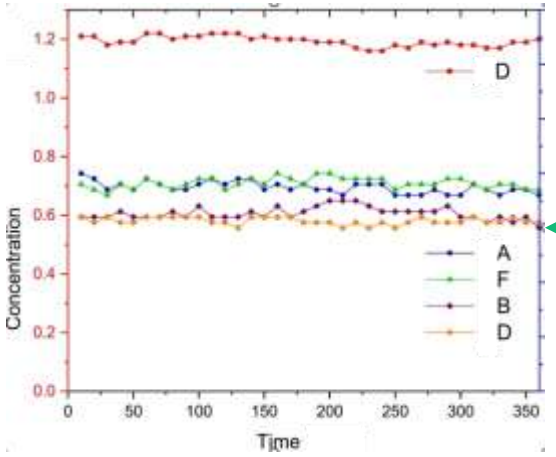


Design of Experiment (DOE)

Model

Machine Learning Models for The Digital Twin

Realtime Monitoring on cloud

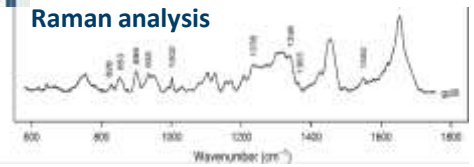


ML Models

Data Platform

Reactor

Raman analysis

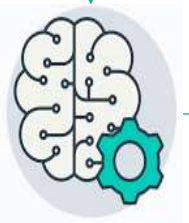


A highly valuable data resource containing 100 batches with 113935 dataset for all available process and Raman spectroscopy measurements from the **University of Manchester** and **University College Londo**n is used for model training and Big Data analytics in the digital twin platform.

[1]Modern day monitoring and control challenges outlined on an industrial-scale benchmark fermentation process, Computers and Chemical Engineering 130 (2019) 106471

Web app for Enduser

Investigated conditions



ML Models

Data Platform

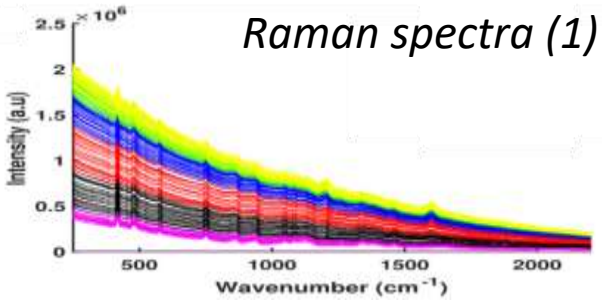
Optimal Targets and Conditions

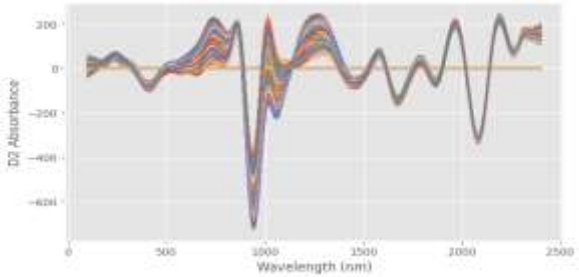
Machine Learning Models for The Digital Twin

PLSr Model

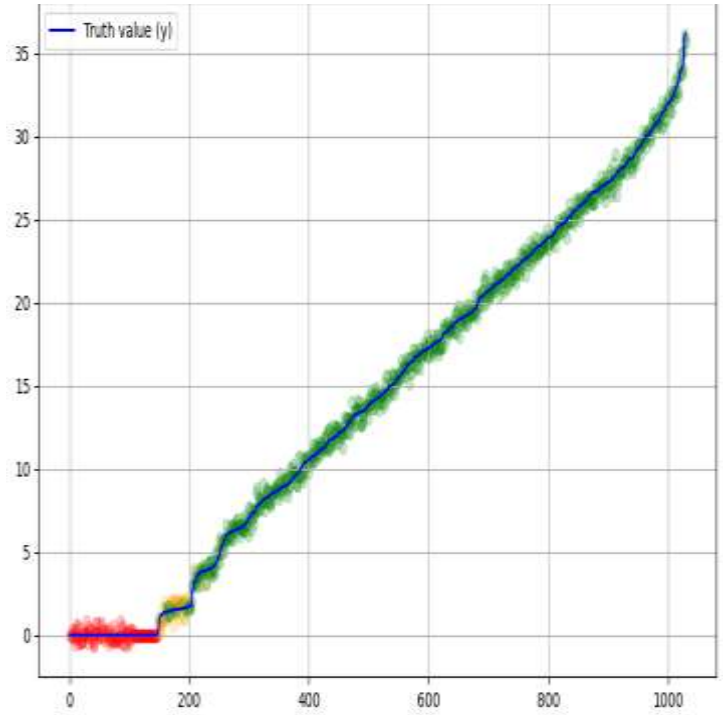
- ✓ *Used to convert Raman spectra into chemical concentrations*
- ✓ *The savgol_filter function in Python is used for smoothing data and eliminating background noise for model development*
- ✓ *2062 datasets applied (50% train - 50% test)*
- ✓ *Achieved R2: 0.997 - MSE calib: 0.228*

Raman spectra (1)





Raman spectra treated by savfol-filter

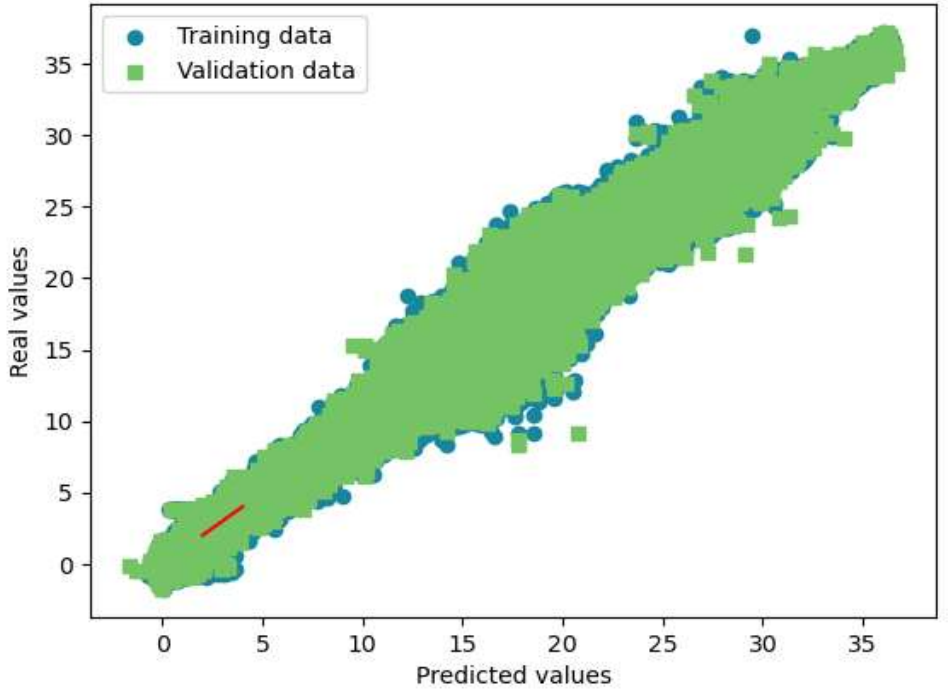


[1]*Modern day monitoring and control challenges outlined on an industrial-scale benchmark fermentation process, Computers and Chemical Engineering 130 (2019) 106471*

Machine Learning Models for The Digital Twin

XGBoost Model

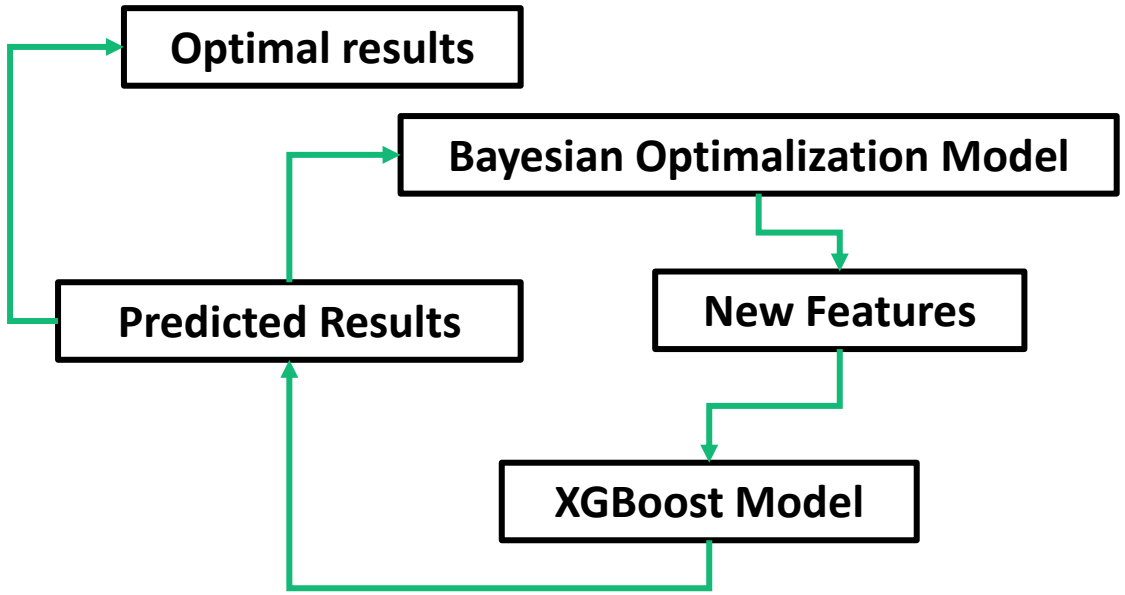
- ✓ *Predicts target concentrations based on operating conditions*
- ✓ *Trained on >100,000 datasets (1)*
- ✓ *80% train / 20% test - Cross validation*
- ✓ *Mean Squared Error Accuracy: 1.854*



[1]*Modern day monitoring and control challenges outlined on an industrial-scale benchmark fermentation process, Computers and Chemical Engineering 130 (2019) 106471*

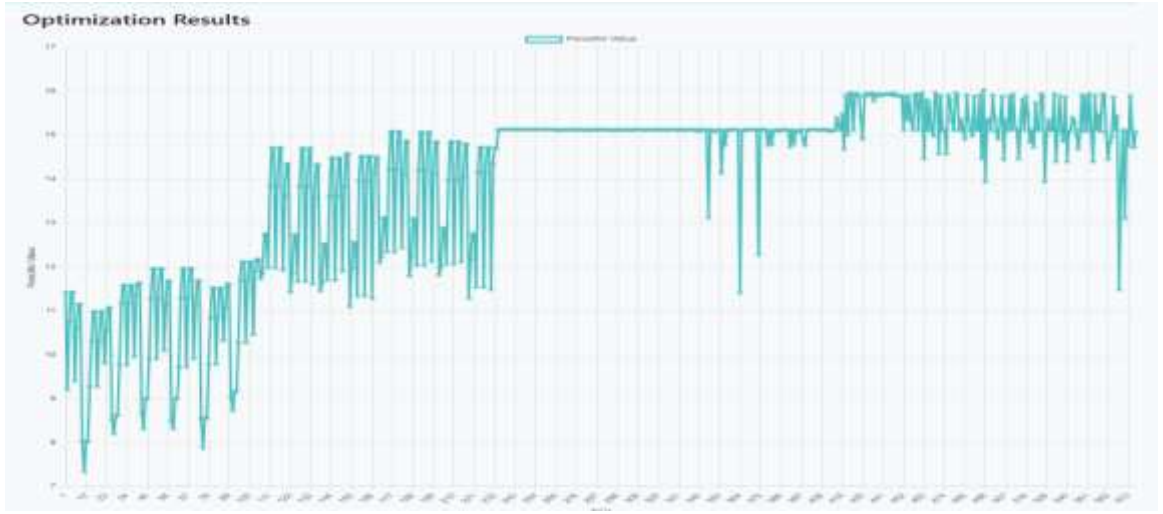
Machine Learning Models for The Digital Twin

The Bayesian Optimization (BO) framework was initialized with 200 random points and further executed with 500 iterations to optimize the target.



👉 *BO performed effectively after the initial 200 random points and showed significant improvement in the optimal values after ~400 iterations.*

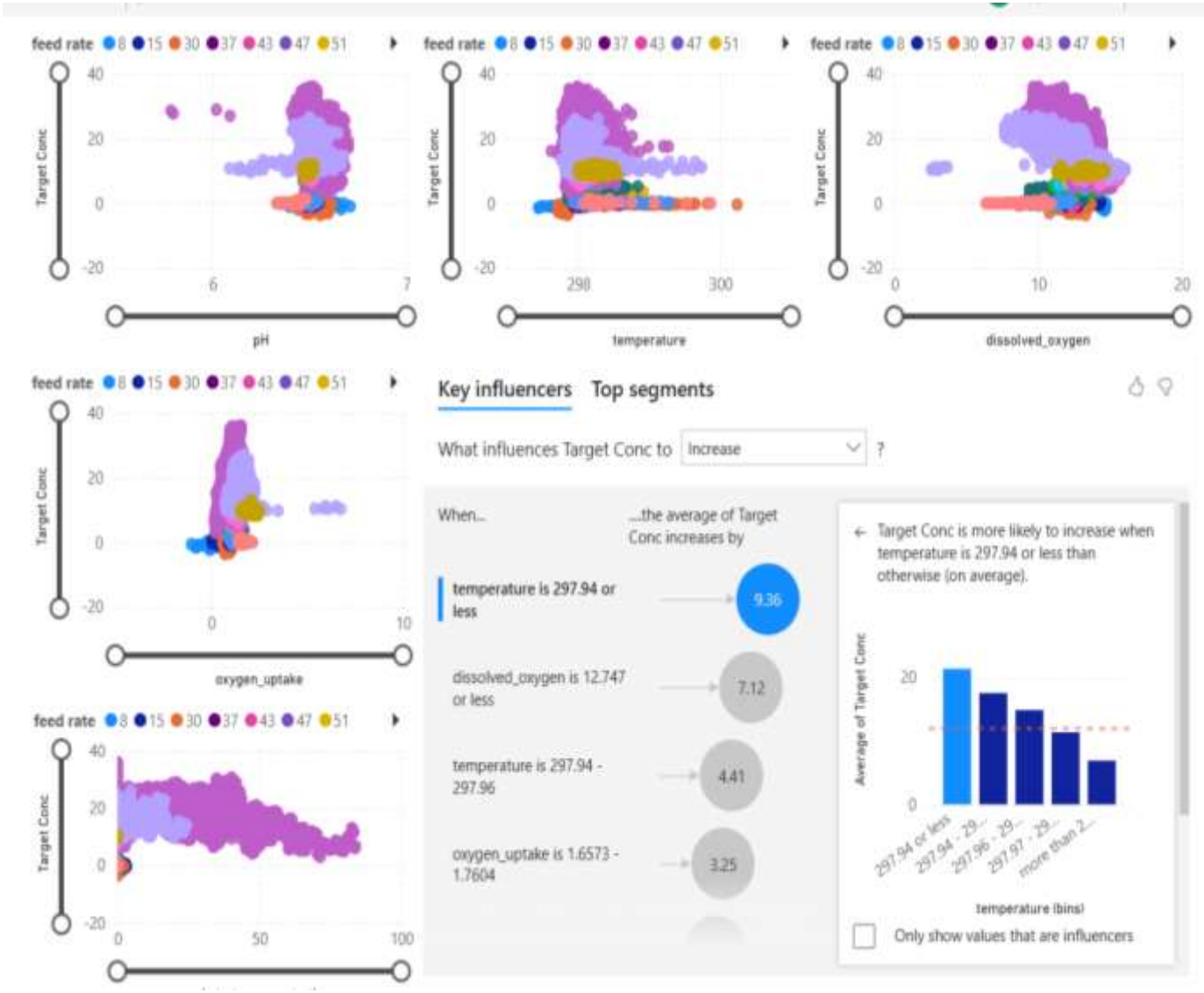
👉 *Supports pilot optimization, reduces experimental costs, and enhances* process *efficiency*



Big Data Analytics for RD

- 113 935 datasets[1] were processed and structured into a project–campaign–test hierarchy. The Power BI dashboard enables fast multi-level queries and real-time updates, supporting trend analysis and factor identification
- *The Power BI dashboard enables fast access to Big Data with multi-level relationships and is automatically updated from the database.*

👉 *It supports the identification of* ***target evolution trends*** *and the* ***factors influencing target performance***



[1]*Modern day monitoring and control challenges outlined on an industrial-scale benchmark fermentation process, Computers and Chemical Engineering 130 (2019) 106471*

Conclusion

- *SINTEF has successfully developed a digital twin platform*
  - ✓ *based on open-source technologies and deployed on Azure Cloud VM.*
  - ✓ *The system integrates ML models, data processing pipelines, and a time-series database, operating effectively for both real-time and batch processing.*
- *The integrated Big Data analytics and optimization models provide strong support for*
  - ✓ *accelerating research*
  - ✓ *enabling faster and more accurate decision-making.*


Thank
you!!!