SINTEF

DATA.ZERO

Edge-based Data Profiling and Repair as a Service for IoT

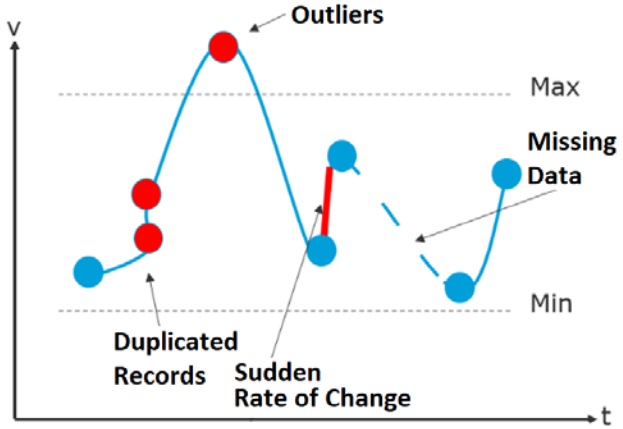
Simeon Tverdal (SINTEF), Arda Goknil (SINTEF), Phu Nguyen (SINTEF), Erik Johannes Husom (SINTEF), Sagar Sen (SINTEF), Jan Ruh (TTTech Computertechnik AG), Francesca Flamigni (TTTech Computertechnik AG)

ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Commission's H2020 Programme under the grant agreement number 958363 (DAT4.Zero).

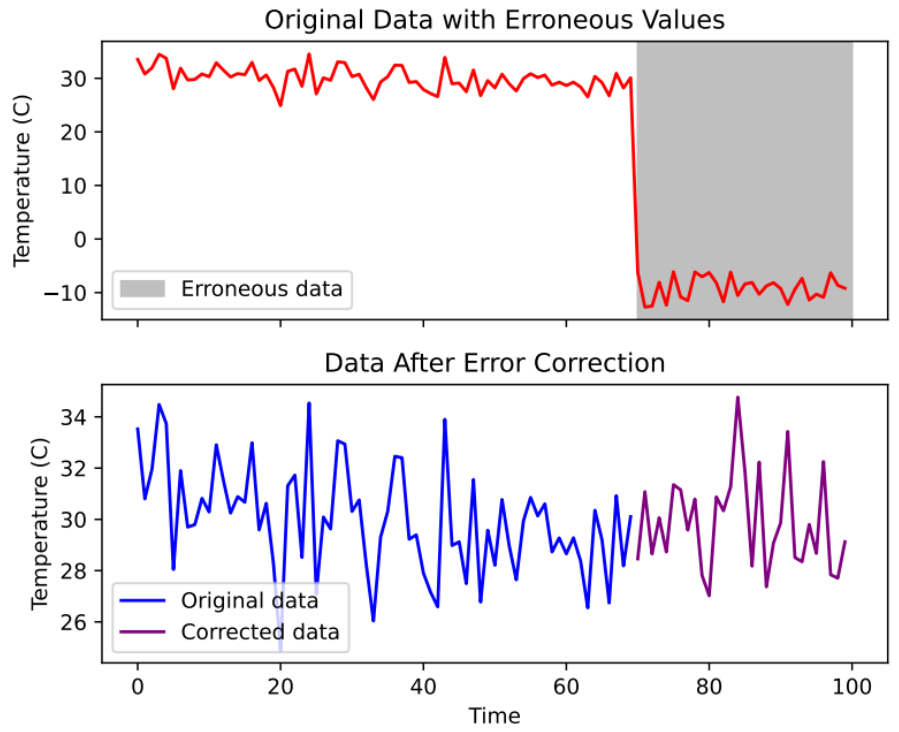
Context: Data Quality for IoT

- The remarkable expansion of IoT results in **vast volumes of data**
- The quality of data collected from IoT devices is often compromised.
  - due to, e.g., sensor inaccuracies, network latency, and environmental conditions
- Data quality issues, including **incomplete or inconsistent data**, can significantly impact the reliability and effectiveness of IoT applications and services



Context: Data Quality for IoT

- Two critical and complementing data quality management techniques for IoT systems are
  - **data profiling** (checking data for data quality requirements)
  - **data repair** (repairing data for detected data quality problems)
- Traditional approaches to data profiling and repair often rely on **centralized architectures**
- This centralized approach poses challenges concerning:
  - network bandwidth
  - scalability
  - privacy
  - security
- **Edge computing emerged as a promising paradigm to address these challenges.**



Related Work

- Considerable research devise and employ **data profiling and repair techniques on the cloud** [1]-[6]

- **Only a few approaches** [7]-[10] detect and repair "corrupt" data at edge/fog devices **near the data source**
  - They have specific constraints that **limit their applicability and generalizability** in IoT applications [11]
  - For instance, Lin et al. [7] require all dependent data computations in the application state history (not always available)

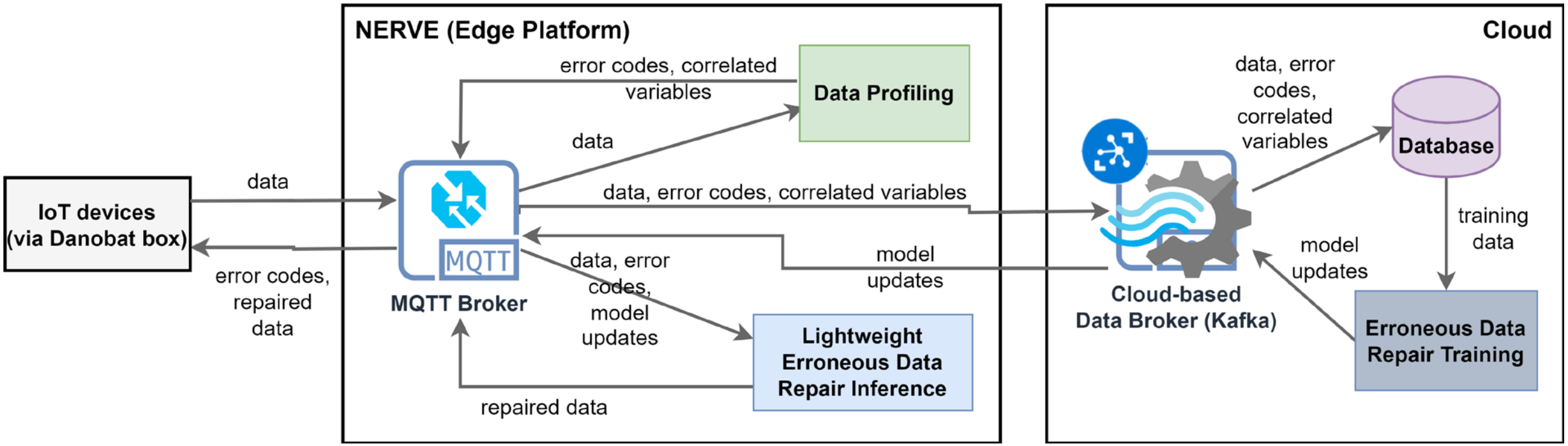
- To overcome such limitations, **Machine Learning** offers a promising solution that can be combined with **existing data profiling techniques**.
  - **Learning correlations among data sources (sensors)**, enabling the substitution of sensors, **prediction of missing values**, and **generation of new data to replace corrupt data**

- **ML-based data repair techniques** can be deployed **at the edge or in the cloud**, leveraging available training data to create containerized repair services for real-time data repair.

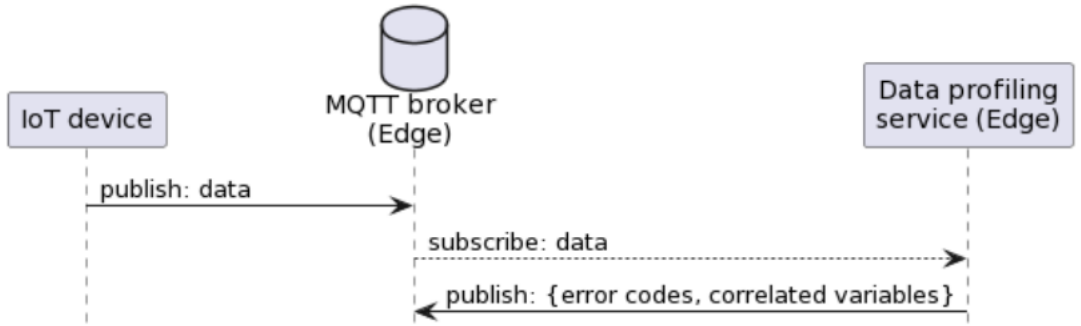
Our Solution: EDPRaaS (Edge-based Data Profiling and Repair as a Service)

- EDPRaaS leverages **pandas profiling [12]** and **Great Expectations [13]** tools for data profiling tasks.
- We utilize **the Nerve Edge platform [14]** as the runtime environment
- The data repair component of EDPRaaS involves **training an ML model on the cloud**
  - then **deployed to the edge for real-time data repair**
- **The ML-based data repair** can effectively handle complex patterns and relationships within the data
  - learn from **existing data patterns** and make predictions or corrections for **erroneous or missing data**

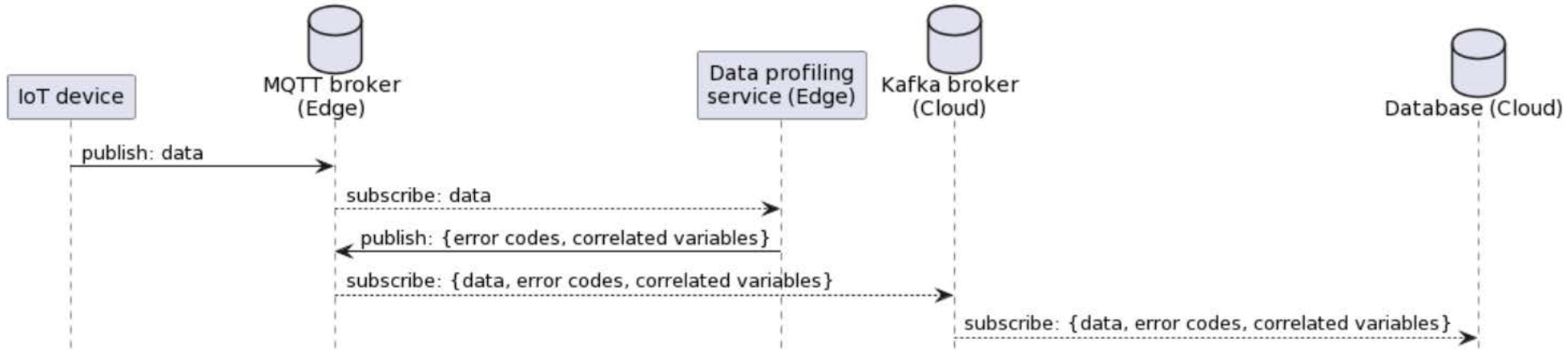
Our Solution: EDPRaaS (Edge-based Data Profiling and Repair as a Service)



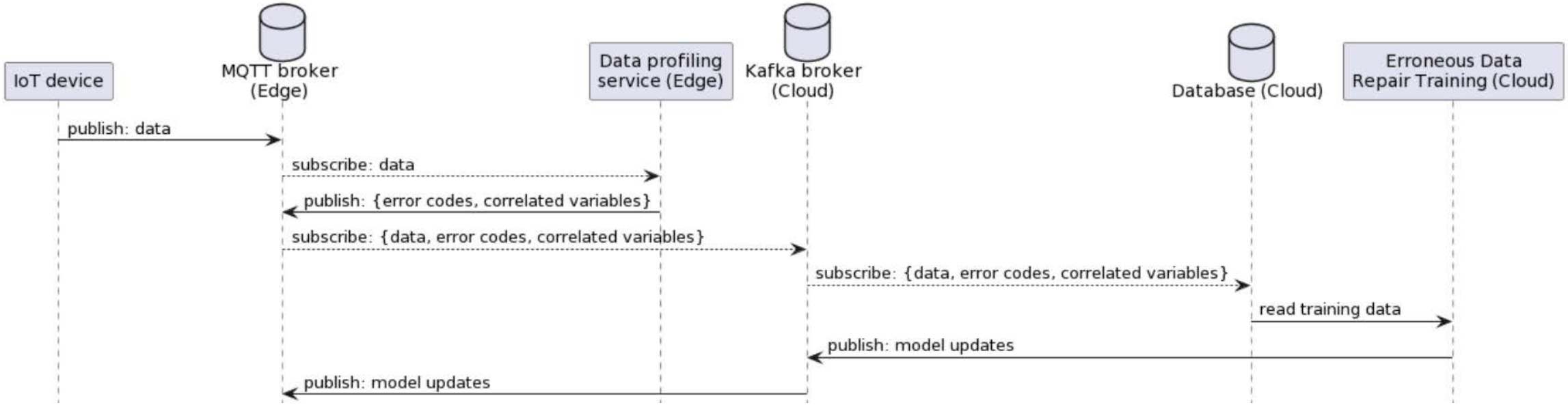
Our Solution: EDPRaaS (Edge-based Data Profiling and Repair as a Service)



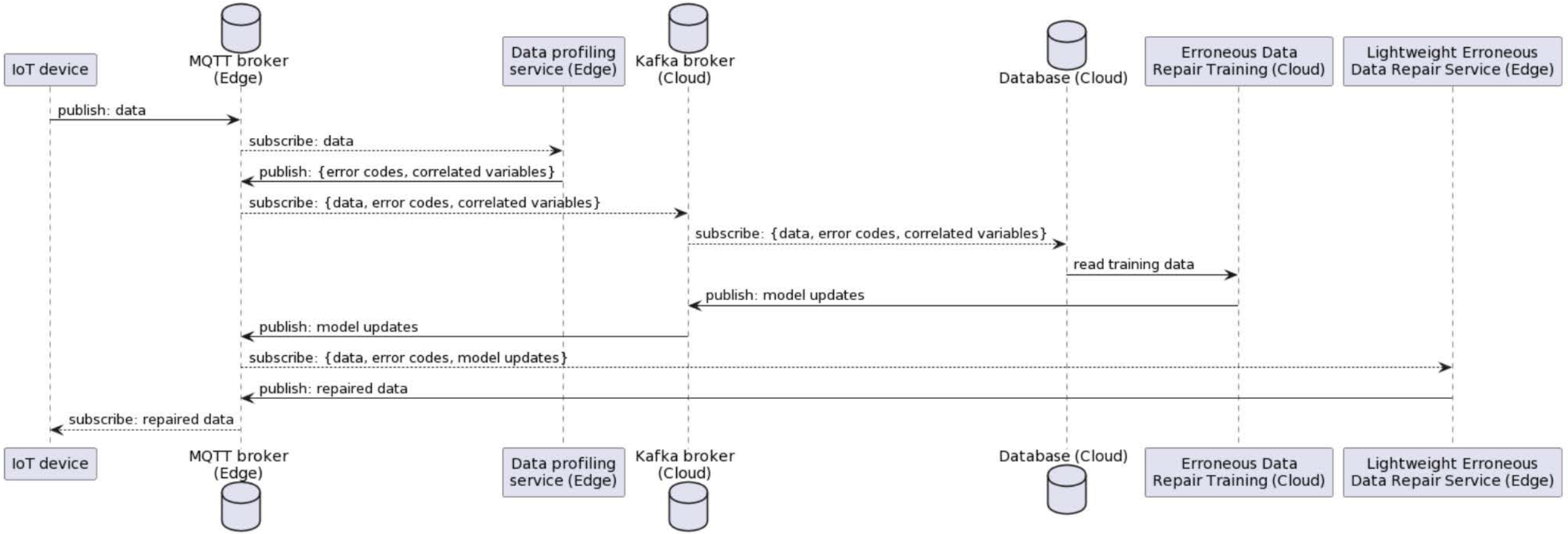
Our Solution: EDPRaaS (Edge-based Data Profiling and Repair as a Service)



Our Solution: EDPRaaS (Edge-based Data Profiling and Repair as a Service)



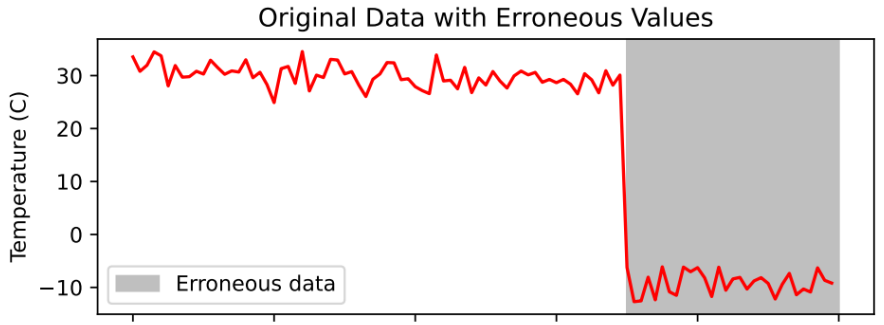
Our Solution: EDPRaaS (Edge-based Data Profiling and Repair as a Service)



EDPRaaS: Data Profiling at the Edge

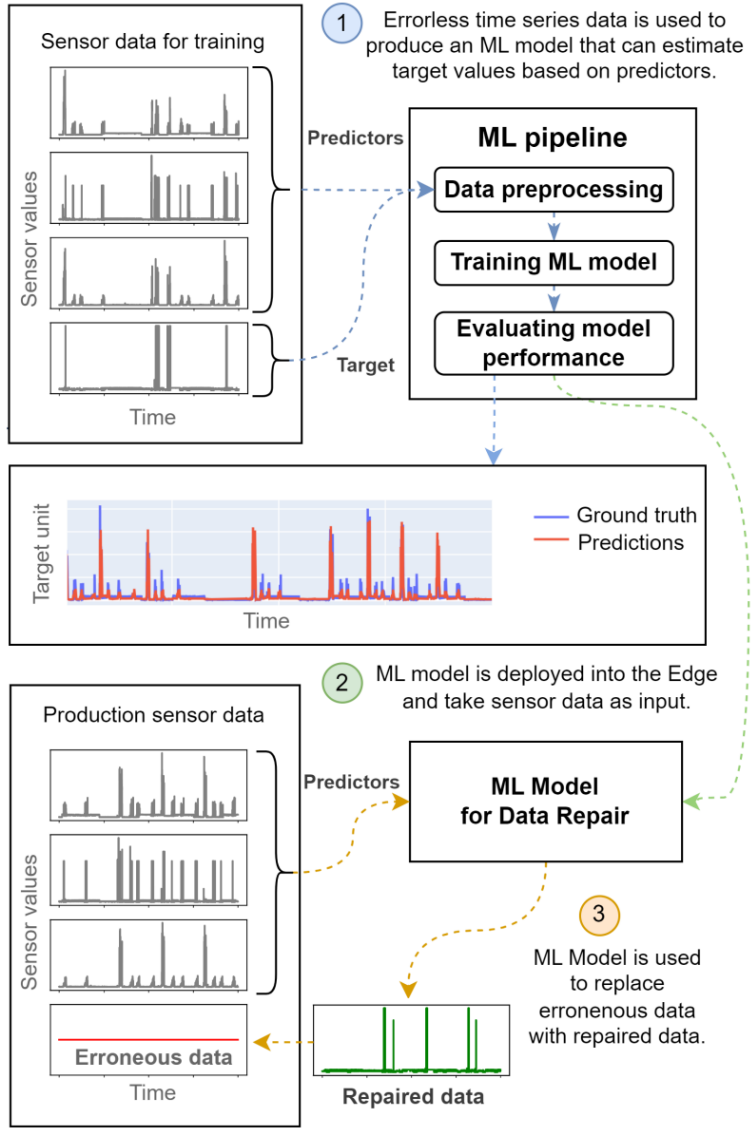
- **Great Expectations** are used to assert the correctness of the data and validate the data quality
- Employing **Great Expectations**, EDPRaaS uncovers data quality issues, including
  - missing values,
  - outliers,
  - duplications.
- **Pandas Profiling** enables
  - exploratory data analysis,
  - general warnings related to possible errors,
  - and finding correlations between the variables in the dataset.

```
{
        "expectation_type": "expect_column_values_to_be_between",
        "kwargs": {
                "column": "Temperature_X",
                "max_value": 40,
                "min_value": 20
        },
        "meta": {}
}
```



EDPRaaS: Data Repair at the Edge

- ML model training is conducted on a cloud platform like Amazon AWS or Microsoft Azure
- **The Erroneous Data Repair Training pipeline** utilizes training data, including error-free data, to train and validate one or more ML models
- These models are trained to **replace erroneous values in variables based on the data profile** generated during data acquisition.



Evaluation – Research Questions

RQ1: *To what extent can EDPRaaS repair erroneous data in a constant data stream?*

RQ2: *Is EDPRaaS capable of timely response to erroneous data?*

RQ3: *How can EDPRaaS be deployed and run on the edge for industrial environments?*

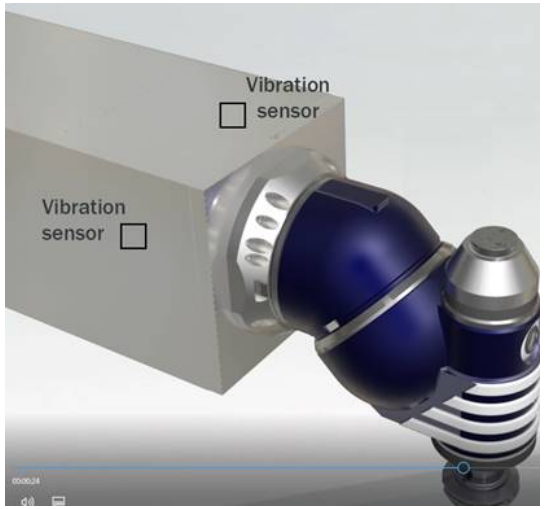
Experiment Design – ML Models

- Our evaluation used two different model configurations
  - **light configuration:** one variable and no feature engineering
  - **heavy configuration:** a lower window size and all variables as input with feature engineering

- We experimented on
  - XGBoost (XBG) and Stochastic Gradient Descent (SGD)
  - Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN)

Experiment Design – Datasets

- **The first dataset (FERSA dataset)** contains accelerometer data collected at a rate of 1 HZ.
- **The second dataset (Automotive dataset)** contains 14 variables collected at 10 HZ from the milling process of a cylinder head's combustion chamber.
- **The third dataset (Aeromec dataset)** contains standard CNC machine data, including temperature measurements and high-frequency accelerometer data.



Experiment Design - Setup

- EDPRaaS is assessed using the Nerve platform and runs on an MFN 100,
  - a Nerve device optimized for use with the Nerve software.

- The Nerve device is designed for harsh industrial environments (-40°C to +70°C).
  - It is based on an Intel Atom x5-E3940/50 CPU and offers 4 GB/8 GB RAM and up to 512 GB SSD storage.

- The MFN 100 offers one I/O port for Ethernet-based fieldbus connectivity, four GbE switch ports, and one SFP port.

RQ1 – Performance of ML-based Data Repair in a Constant Data Stream?

Model	Fersa Dataset					
	Light Configuration			Heavy Configuration		
	Late.	R^2	MSE	Late.	R^2	MSE
XGB	4.38	0.791	4.04e-09	5.28	0.8741	2.34e-09
SGD	4.35					
RNN	5.46					
CNN	4.70					

Model	Automotive Dataset					
	Light Configuration			Heavy Configuration		
	Late.	R^2	MSE	Late.	R^2	MSE
XGB						
SGD						
RNN						
CNN						

Model	Aeromec Dataset					
	Light Configuration			Heavy Configuration		
	Late.	R^2	MSE	Late.	R^2	MSE
XGB	5.73	0.82	0.003	6.28	0.99	0.0002
SGD	4.96	0.25	0.011	6.72	-8.48e+20	1.26e+19
RNN	6.13	0.03	0.015	7.94	0.80	0.0029
CNN	5.47	0.82				

Answer to RQ1: Selecting the ML architecture and configuration for EDPRaaS should consider dataset characteristics and the edge use case. Balancing model complexity with performance risk is crucial. For scenarios where reliability and consistency are crucial, adopting simpler ML architectures like XGBoost, which demonstrate relatively high R^2 scores across datasets, may be preferred. Understanding trade-offs allows informed decision-making to choose the most suitable ML solutions for efficient and accurate data repair in EDPRaaS.

RQ2 – Timely response of EDPRaaS to Erreneous data?

Model	Fersa Dataset					
	Light Configuration			Heavy Configuration		
	Late.	R^2	MSE	Late.	R^2	MSE
XGB	4.38	0.791	4.04e-09	5.28	0.8741	2.34e-09
SGD	4.35	0.693	5.94e-9	4.49	0.7515	4.63e-9
RNN						
CNN						

Model	Aeromec Dataset					
	Light Configuration			Heavy Configuration		
	Late.	R^2	MSE	Late.	R^2	MSE
XGB	5.73	0.82	0.003	6.28	0.99	0.0002
SGD	4.96	0.25	0.011	6.72	-8.48e+20	1.26e+19
RNN	6.13					
CNN	5.47					

Answer to RQ2: EDPRaaS efficiently repairs erroneous data using both simple and complex ML models at the edge. Our findings indicate that deploying resource-intensive deep learning models is feasible when their training occurs in the cloud.

RQ3 – EDPRaaS deployed and run on the edge for industrial environments?

- Data versioning and dependency management
- Deployment infrastructure
- Monitoring and logging
- Integration of models and components

Answer to RQ3: EDPRaaS is successfully deployed on the edge for industrial environments. It leverages DVC for data versioning and dependency management and employs Docker containers for streamlined deployment. Monitoring and logging capabilities are enabled through DVC and the Nerve software infrastructure. Integration of ML models and components, facilitated by ML frameworks and Docker, ensures seamless operation and efficient online data repair at the edge.

Conclusion

- *EDPRaaS (Edge-based Data Profiling and Repair as a Service)*
  - *a novel and efficient approach for data quality enhancement in IoT environments*

- EDPRaaS addresses the challenges of
  - real-time decision-making,
  - bandwidth efficiency,
  - resource constraints, and
  - online operation.

- The integration of ML in the data repair component enhances the accuracy and adaptability of the repair process,
  - ensuring continuous data quality improvement

ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Commission's H2020 Programme under the grant agreement number 958363 (DAT4.Zero).

References (1)

[1] Dominik Flick, Sebastian Gellrich, Marc-André Filz, Li Ji, Sebastian Thiede, and Christoph Herrmann. 2019. Conceptual Framework for manufacturing data preprocessing of diverse input sources. In INDIN'19. 1041–1046.

[2] Mohammad Ayoub Khan and Fahad Algarni. 2020. A Healthcare Monitoring System for the Diagnosis of Heart Disease in the IoMT Cloud Environment Using MSSO-ANFIS. IEEE Access 8 (2020), 122259–122269.

[3] Tianxiang Kong, Tianliang Hu, Tingting Zhou, and Yingxin Ye. 2021. Data Construction Method for the Applications of Workshop Digital Twin System. Journal of Manufacturing Systems 58 (2021), 323–328.

[4] Sagar Sen, Erik Johannes Husom, Arda Goknil, Dimitra Politaki, Simeon Tverdal, Phu Nguyen, and Nicolas Jourdan. 2023. Virtual sensors for erroneous data repair in manufacturing a machine learning pipeline. Computers in Industry 149 (2023), 103917.

[5] Chang Wang, Yongxin Zhu, Weiwei Shi, Victor Chang, P. Vijayakumar, Bin Liu, Yishu Mao, Jiabao Wang, and Yiping Fan. 2018. A Dependable Time Series Analytic Framework for Cyber-Physical Systems of IoT-Based Smart Grid. ACM Transactions on Cyber-Physical Systems 3, 1 (2018), 18.

[6] Sholom M. Weiss, Amit Dhurandhar, and Robert J. Baseman. 2013. Improving Quality Control by Early Prediction of Manufacturing Outcomes. In KDD'13. 1258–1266

[7] Wei-Tsung Lin, Fatih Bakir, Chandra Krintz, Rich Wolski, and Markus Mock. 2019. Data Repair for Distributed, Event-Based IoT Applications. In DEBS'19. 139–150.

References (2)

[8] Nwamaka U. Okafor, Yahia Alghorani, and Declan T. Delaney. 2020. Improving Data Quality of Low-cost IoT Sensors in Environmental Monitoring Networks Using Data Fusion and Machine Learning Approach. ICT Express 6, 3 (2020), 220–228.

[9] Luke Russell, Felix Kwamena, and Rafik Goubran. 2019. Towards Reliable IoT: Fog-Based AI Sensor Validation. In IEEE Cloud Summit. 37–44.

[10] Sunny Sanyal and Puning Zhang. 2018. Improving Quality of Data: IoT Data Aggregation Using Device to Device Communications. IEEE Access 6 (2018), 67830–67840.

[11] Arda Goknil, Phu Nguyen, Sagar Sen, Dimitra Politaki, Harris Niavis, Karl John Pedersen, Abdillah Suyuthi, Abhilash Anand, and Amina Ziegenbein. 2023. A Systematic Review of Data Quality in CPS and IoT for Industry 4.0. ACM Comput. Surv. 55, 14s, Article 327 (jul 2023)

[12] pandas-profiling. [n. d.]. https://pypi.org/project/pandas-profiling/. Visited in 2023.

[13] Great Expectations. [n. d.]. https://greatexpectations.io/. Visited in 2023.

[14] Nerve Platform. [n. d.]. https://docs.nerve.cloud/. Visited in 2023.

SINTEF

Technology for a better society