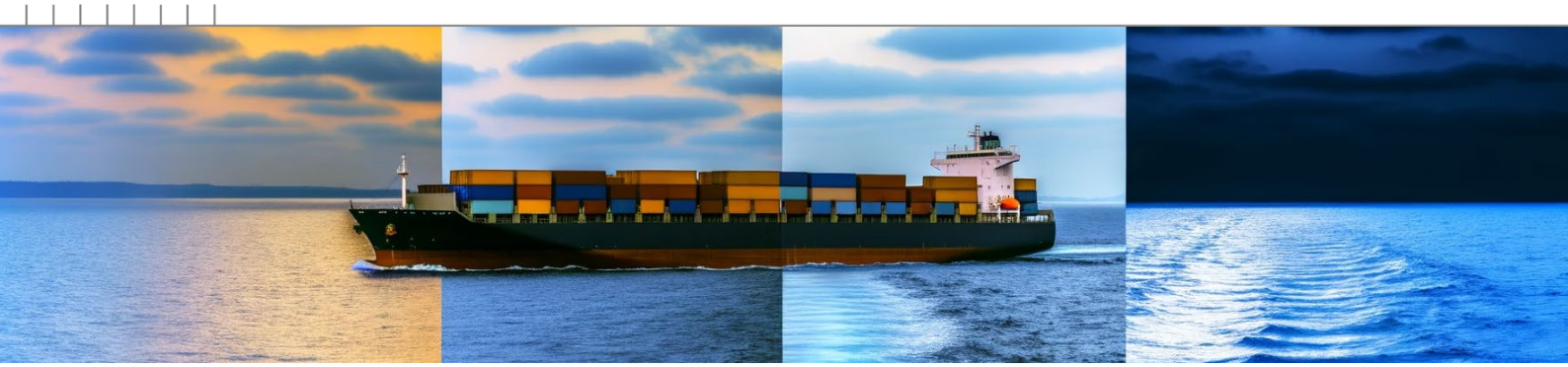


Training a machine learning model without sharing data using Federated Learning

Data accessibility is a key concern in maritime operations. While machine learning can be limited when used locally due to small datasets and potential biases, there are often legal and technological hurdles that make it difficult to centralize this data on a single server, which could otherwise help generate valuable insights for the industry. This has led to the emergence of Federated Learning, a key technology for developing high-quality predictive models without transferring sensitive data.

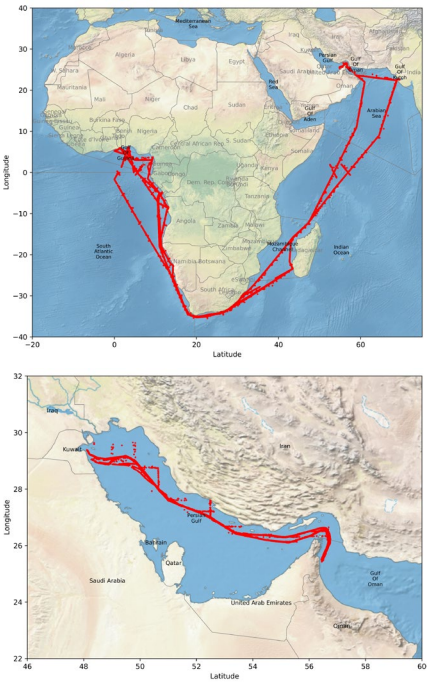


Figure 1: Location of recordings in 2012 (top) and 2014 (bottom)

In the rapidly evolving landscape of machine learning, data privacy and security are predominant concerns. Traditional centralized approaches to machine learning involve aggregating all data into a single location where models are trained. However, this method poses significant privacy risks and often faces regulatory challenges, especially in sensitive industries such as healthcare, finance, and maritime operations.

Federated Learning leverages the collective intelligence of multiple decentralized devices or servers, allowing them to collaboratively train a shared model while keeping the data localized. This innovative approach, popularized by Google, not only mitigates privacy risks but also enhances the robustness and accuracy of machine learning models by tapping into diverse data sources. By maintaining data privacy and security, Federated Learning aligns with regulatory requirements and standards, making it an attractive option for industries that manage sensitive information.

After a theoretical and technical overview of Federated Learning we conducted, within CRS [2], a comparative analysis through the case of a bulk carrier sailing along the African and Persian coasts. This dataset includes thousands of measurements from 2012 to 2014, which were used in training our model. In 2012, the ship navigated the Atlantic and Indian Oceans, while in 2014, it was in the Persian Gulf, resulting in operational environment's variability

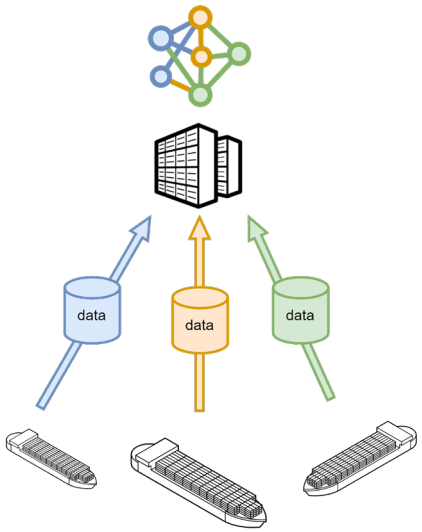


Figure 2: In a typical configuration, data sets are stored and aggregated on a server. A model can then be trained on this centralized dataset.

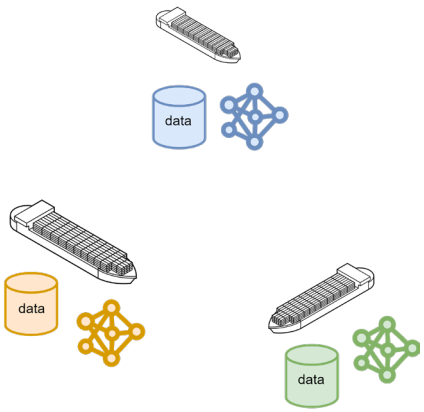


Figure 3: If data exchange is not possible, each customer can train its own model, but with no guarantee that it will work on unseen data.

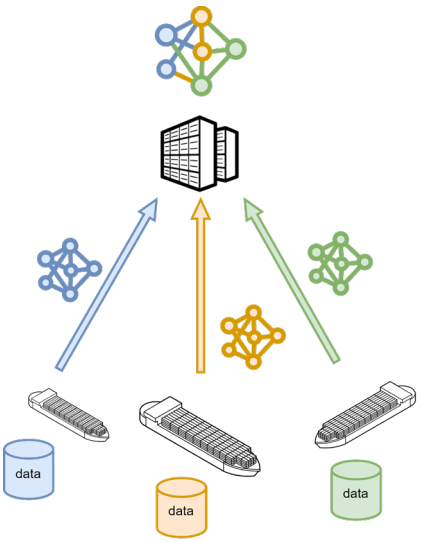


Figure 4: Alternatively, federated learning allows to learn a common, more general model by iteratively exchanging only the model's characteristics (its weights, for example).

If all the data points can be used, without worrying about privacy or other technical issues, we end up with a typical machine-learning configuration, as illustrated in the Figure 2. The model is consequently trained with a wide range of situations in its training set, and we find that it works well across all our test data. This model served as our baseline for the study.

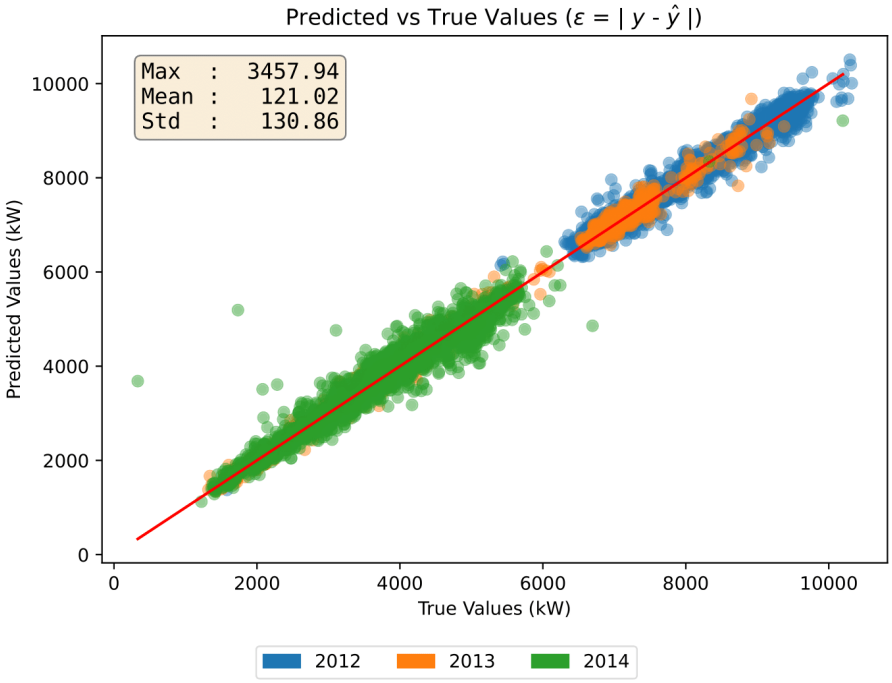


Figure 5: Predictions of the centralized model on test data from each year.

What if each dataset belongs to different customers and they can't or don't want to share it, as shown in Figure 3? In that case, models can still be trained individually. These isolated models will probably perform well on their respective datasets, but may not work under other conditions, as suggested by Figure 6.

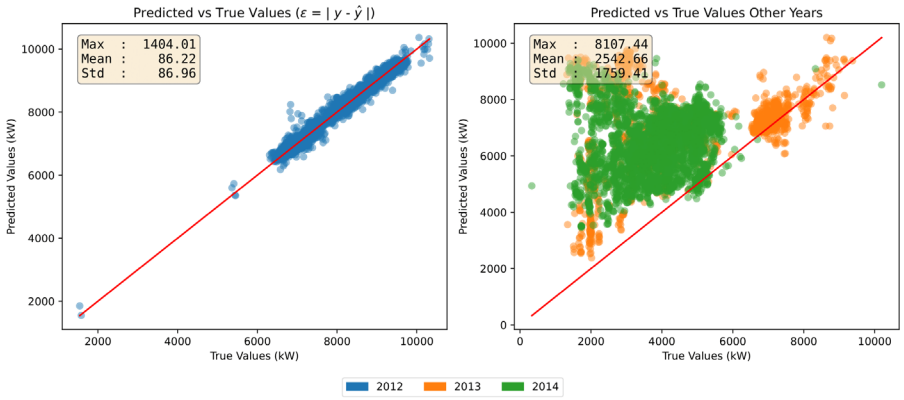


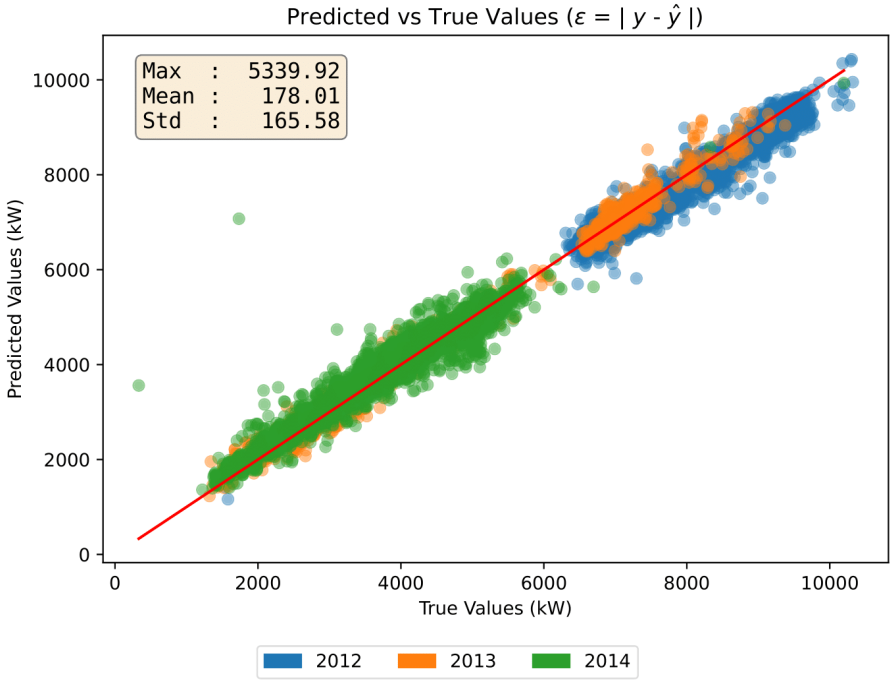
Figure 6: Predictions of a model trained only with data from 2012. When encountering unseen conditions from other years, the quality of the predictions degrades.

Our use case is a blatant example of such non-generalization. The model trained on 2012 data (in the Atlantic and Indian oceans) contains almost no records where the ship's power is below 6000 kW while this is the typical range of our recordings in the Persian Gulf. This could be explained by a number of factors, including traffic density, environmental conditions, as well as international and local regulations.

Federated learning will not solve all the problems associated with proprietary data or the transmission of such data in inadequate environments. Heterogeneous data and devices represent significant challenges for the researchers working on them, not to mention the interpretability of a model whose entire training dataset is not accessible. Nonetheless, the collaborative nature of this decentralized approach, aligned with the core objectives of CRS and MARIN, opens the horizon to exciting new projects and solutions in a world where more and more data is being produced yet not used.

For more information contact MARIN:
Gaspard Ducamp
T + 31 317 49 37 06
E g.ducamp@marin.nl

All hope of finding a high-performance, general-purpose model without directly exchanging proprietary data is not lost. By simulating a federated context on Marclus 5 [3] using the Flower package [4], we have succeeded in training a model without data exchange that gives results close to the centralized model. While this model is certainly not perfect, it does illustrate the value of such an approach.



Through the implementation of federated learning mechanisms, we have demonstrated through a well-known example that despite the disparate nature of maritime data, it is possible to learn meaningful and actionable information without compromising data confidentiality or integrity. This study reaffirms the hypothesis that federated learning can serve as a robust framework for effectively harnessing disparate data sources, for example to facilitate a comprehensive and nuanced understanding of ship performance under varied operational conditions.

The results derived from this study not only highlight the potential for substantial improvements in predictive accuracy and operational efficiency, but also mark a major step towards a more interconnected and collaborative approach to maritime analytics.

Reference:

[1] Federated Learning: Building better products with on-device data and privacy by default. An online comic from Google AI.

[2] Cooperative Research Ships | MARIN.

[3] Facilities & Tools | MARIN.

[4] Beutel, D. J., Topal, T., Mathur, A., Qiu, X., Fernandez-Marques, J., Gao, Y., Sani, L., Li, K.H., Parcollet, T., Gusmao, P.P.B. de and Lane, N.D.: *"Flower: A Friendly Federated Learning Research Framework"*, 2020.