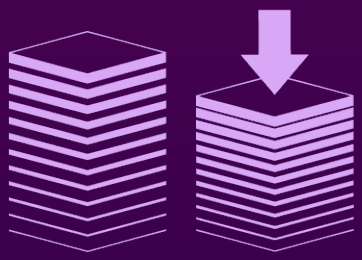


Lightscline

Data Reduction AI

# Analyzing the 10% important data from Terabytes of DFOS data

August 2024

Lightscline, State College, PA, USA

## About Lightscline

Lightscline's lightweight AI solves the foundational problem of efficiently analyzing Terabytes of Distributed Fiber Optics Sensing (DFOS) data being generated from applications like urban infrastructure monitoring (city traffic, foot traffic), oil & gas exploration and monitoring, and geosciences. Using Lightscline AI's 4 lines of code that can be setup within 10 minutes, customers can reduce 90% of their sensor data infra & human time and costs by selectively focusing on just 10% of the raw data. By exploiting the redundancy of real-world data, Lightscline AI makes real-time predictions using just 10% important data.

## Executive Summary

In this whitepaper, we show how Lightscline's product can work with DFOS data, and how Lightscline AI can help in generating actionable insights using just 10% of the raw data. We analyze DAS datasets used to monitor subsurface environment changes and detect movement of cars, traffic patterns, and walk detection. We include (i) Data Labeling, (ii) Visualization, (iii) Model training approaches to understand how Lightscline's multi-channel analysis can be used to make the labelling & analysis significantly faster.

## Applications Overview

Lightscline's lightweight AI unlocks several capabilities across space, aerial, terrestrial, and underwater applications which are currently unobtainable. Some applications include:

1. Quickly analyze 200+ hours of acoustic data for anomaly detection and ATR using just 10% important data
2. Order-of-magnitude reduction in the amount of training data without degradation to identification performance (Pid)
3. Prioritize training data by selectively focusing on the 10% important data
4. Enable transfer learning on the edge to quickly train for complementary tasks
5. Real-time inference on 15+ classes of data for on-board Human Activity Recognition (HAR)
6. On-board satellite-based AIS signal validation using 10% of the raw RF data

In this whitepaper, we will focus on DFOS data with applications like:

1. Leak detection & integrity monitoring in oil & gas applications
2. Walk / car / traffic detection using dark cables in urban environments
3. Natural phenomena and seismic activity monitoring
4. Detect micro-seismic events

## Problem: Huge volumes of DFOS data

We analyze a public DFOS dataset available here. The parent folder contains several sub-folders and files in .tdms format, which contain the DFOS data collected over several months of experimentation.

The following figure shows the data files for April 4, 2019, from a public DAS dataset in State College, PA, USA. Data files are in .tdms format and can be sorted according to specific dates and times for visualization in a preferred python IDE.

Dataset size:
Rate of data generation = ~142MB per minute (with just 500Hz sampling frequency)
Size of the data for April alone is greater than 4.5 TB.

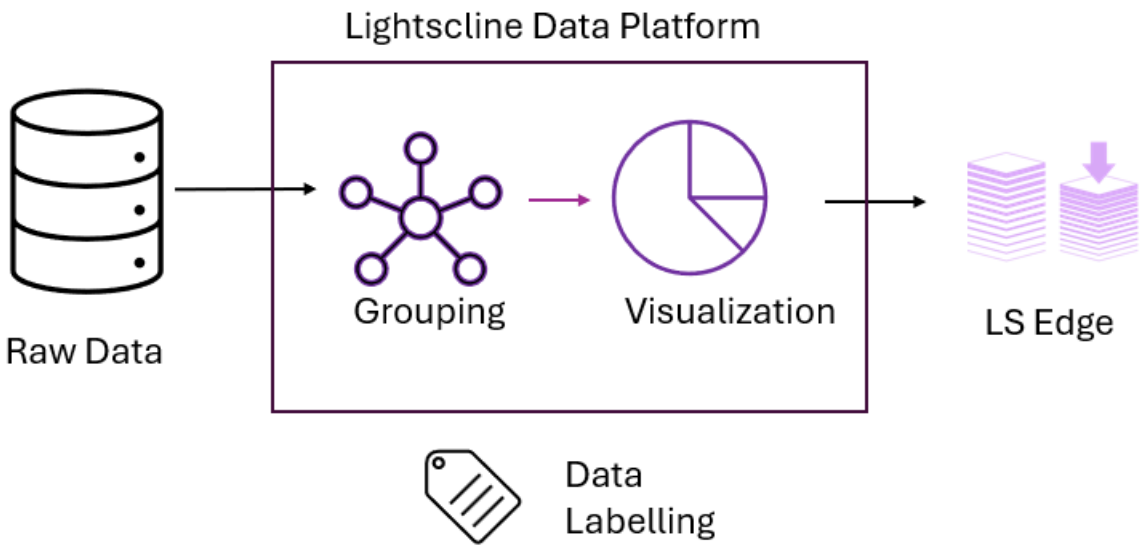
```
=====================================================
Content of ZIP archive  apr-001.zip
=====================================================

Archive:  apr-001.zip
  Length      Date    Time    Name
---------  ---------- -----   ----
        0  05-05-2021 20:30   apr/
149886976  08-22-2019 09:43   apr/PSUDAS_UTC_20190404_194804.509.tdms
149886976  08-22-2019 09:44   apr/PSUDAS_UTC_20190404_194904.509.tdms
149886976  08-22-2019 09:44   apr/PSUDAS_UTC_20190404_195004.509.tdms
149886976  08-22-2019 09:44   apr/PSUDAS_UTC_20190404_195104.509.tdms
149886976  08-22-2019 09:44   apr/PSUDAS_UTC_20190404_195204.509.tdms
149886976  08-22-2019 09:44   apr/PSUDAS_UTC_20190404_195304.509.tdms
149886976  08-22-2019 09:44   apr/PSUDAS_UTC_20190404_195404.509.tdms
149886976  08-22-2019 09:44   apr/PSUDAS_UTC_20190404_195504.509.tdms
149886976  08-22-2019 09:44   apr/PSUDAS_UTC_20190404_195604.509.tdms
149886976  08-22-2019 09:44   apr/PSUDAS_UTC_20190404_195704.509.tdms
149886976  08-22-2019 09:44   apr/PSUDAS_UTC_20190404_195804.509.tdms
149886976  08-22-2019 09:44   apr/PSUDAS_UTC_20190404_195904.509.tdms
149886976  08-22-2019 09:44   apr/PSUDAS_UTC_20190404_200004.509.tdms
149886976  08-22-2019 09:44   apr/PSUDAS_UTC_20190404_200104.509.tdms
149886976  08-22-2019 09:44   apr/PSUDAS_UTC_20190404_200204.509.tdms
149886976  08-22-2019 09:44   apr/PSUDAS_UTC_20190404_200304.509.tdms
149886976  08-22-2019 09:44   apr/PSUDAS_UTC_20190404_200404.509.tdms
149886976  08-22-2019 09:44   apr/PSUDAS_UTC_20190404_200504.509.tdms
149886976  08-22-2019 09:44   apr/PSUDAS_UTC_20190404_200604.509.tdms
149886976  08-22-2019 09:44   apr/PSUDAS_UTC_20190404_200704.509.tdms
```

This voluminous data is challenging to label, process and visualize. Moreover, ML models on this data requires compute heavy hardware, making it impractical to train and deploy on the edge. Lightscline's Data Platform (LDP) helps to quickly label and visualize the data. Next, Lightscline's Edge AI helps in faster training of models and reduces the infrastructure costs.

## 1. Lightscline Data Platform

1.1 Data Grouping:

Lightscline Data Platform groups the data before humans start labeling it. Users can quickly visualize the grouped data using the LDP. By manually labeling a data window from a group, LDP can be used for automatically labeling the entire group. This enabled a Fortune 500 company to manually label only 100 windows instead of the original 10 Million windows – creating a 100,000x efficient labelling process! Grouping similar data saves time & costs and ensures high-quality labeled data.




1.2 Data Visualization:

Data Platform provides data visualization, which is a useful prior for applying any conventional analysis techniques. Visualization is needed to label one member of the group.

Below is a basic but powerful example of information gleaned by mere data visualization.

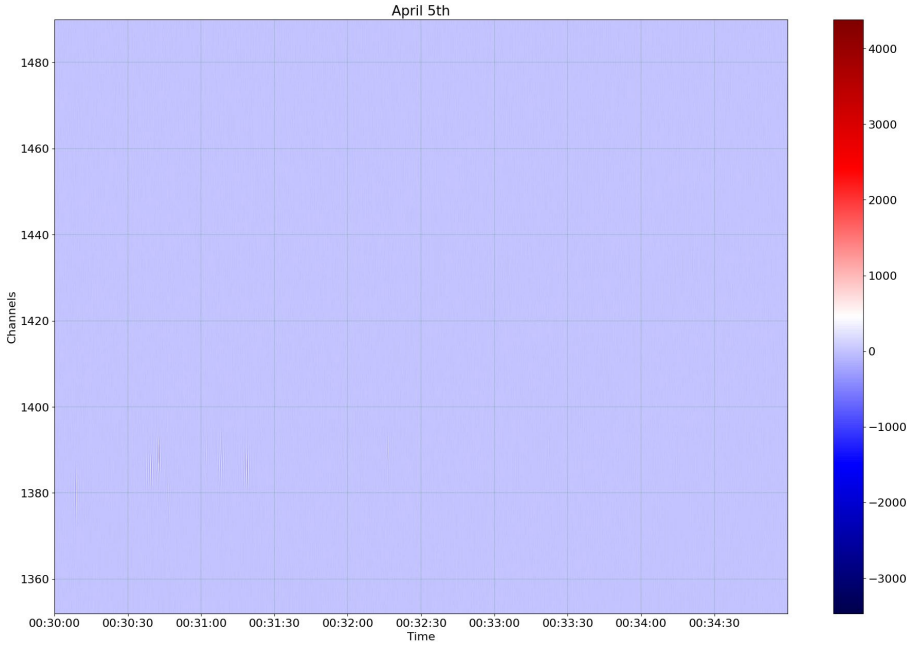
The following channels are set up at State College, PA in the downtown area. State College is a college town with low traffic footprint. This example highlights the power of visualization combined with context to draw conclusions.


Based on the feature seen in the above plot, we can make the following inference:

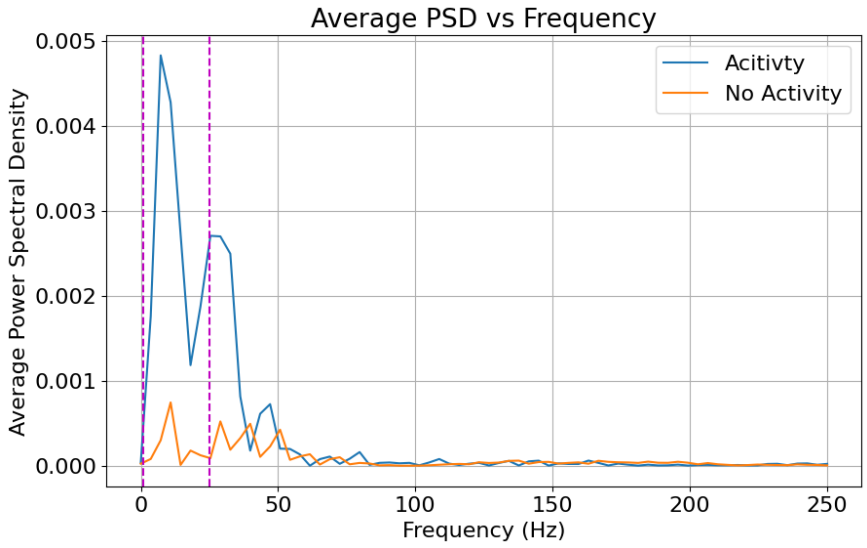
Channels spacing = 2m
Activity can be seen for ~65s
Activity seen from channels = 1460-1400 = 60
Velocity = 120m/65s = 1.85mph
The detected activity is likely walking.

Now all the similar data can be grouped together automatically and labeled “Activity”. This significantly reduces the labeling time.

The following example shows that there is no activity during 12:30-12:35 AM.




All this data can be grouped together and labeled as “No Activity” to create labeled data.

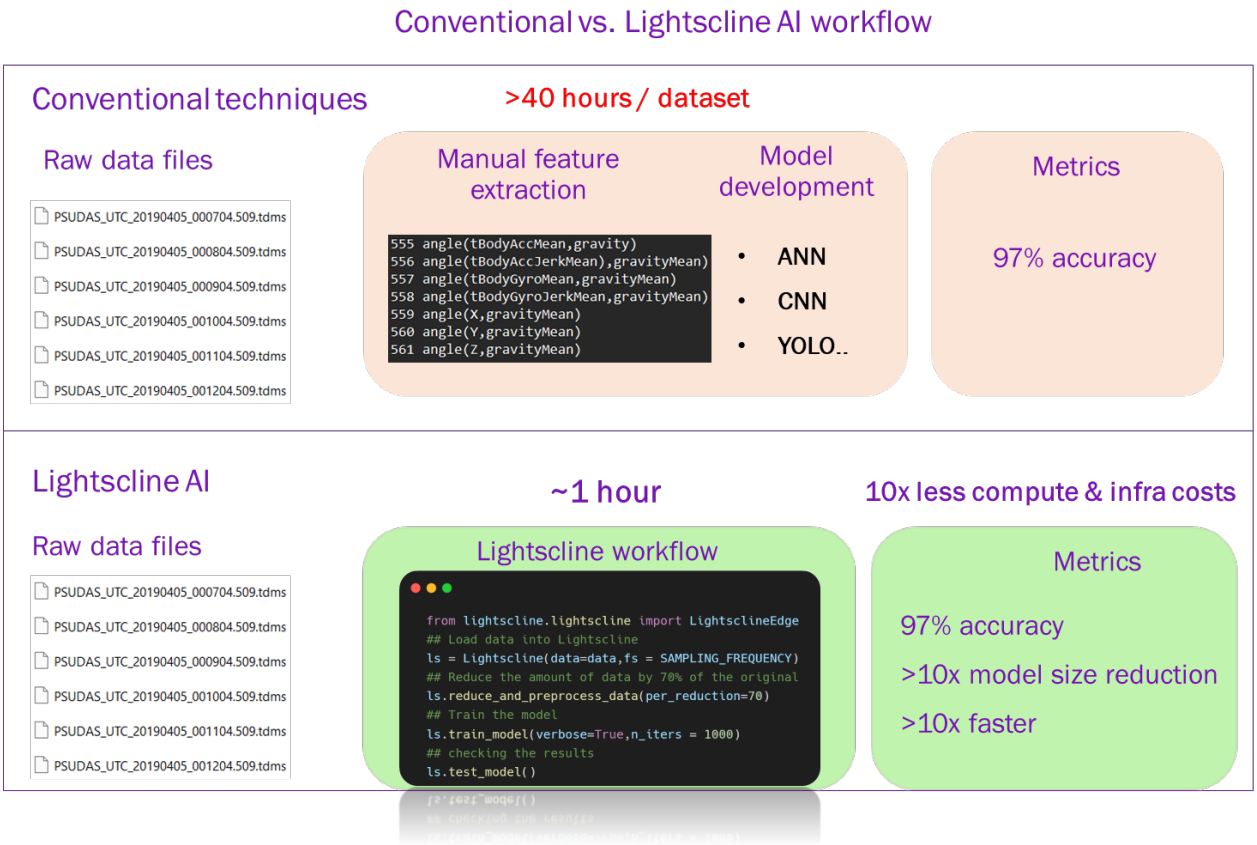



In this graph the average PSD has higher energy in 1-25 Hz frequency range. 1-25 Hz frequency range is generally where footsteps & vehicles are detected.

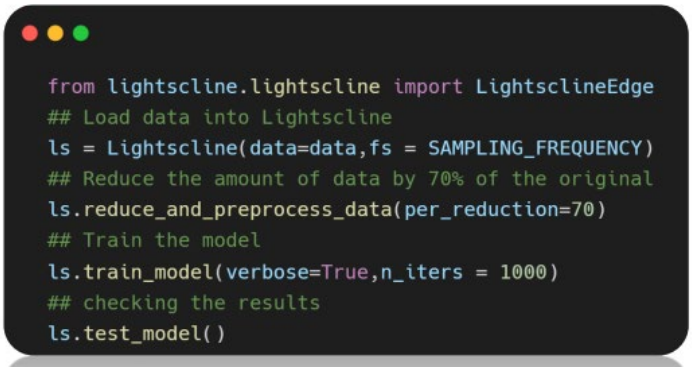
## 2. Lightscline Edge AI

After data labeling, the next steps typically involve either manual feature extraction or training the model on the full raw dataset. While feature extraction can result in smaller models, it is time-consuming because data scientists must manually select the features to be extracted for each dataset. On the other hand, training on the full raw dataset can achieve better accuracy but leads to significantly larger model sizes. By leveraging the redundancy of real-world sensor data, Lightscline Edge automatically identifies a small fraction of important data for different downstream tasks. This approach eliminates the need for manual feature extraction and results in much smaller models, saving both time and costs.

The diagram below illustrates how Lightscline Edge achieves these efficiencies.



## Lightscline AI



```
from lightscline.lightscline import LightsclineEdge
## Load data into Lightscline
ls = Lightscline(data=data,fs = SAMPLING_FREQUENCY)
## Reduce the amount of data by 70% of the original
ls.reduce_and_preprocess_data(per_reduction=70)
## Train the model
ls.train_model(verbose=True,n_iters = 1000)
## checking the results
ls.test_model()
```

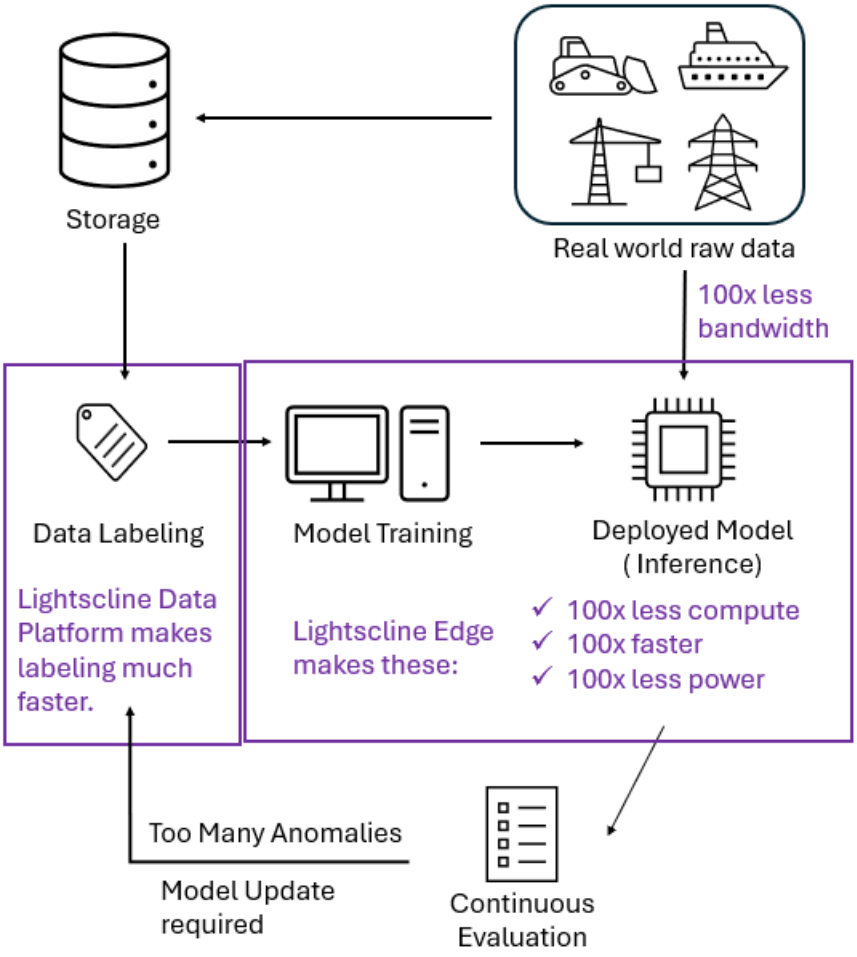
- 4 lines of code to get started
- Setup within 10 mins
- No data sharing required

Several dual-use customers have used Lightscline edge AI to see the benefits of Lightscline edge themselves. Users can get started with Lightscline AI using just 4 lines of code that can be setup within 10 minutes, without any need for external data sharing. This product can be run on the customer's cloud environment or on-prem. This product can also handle multi modal data coming from different sensors at different sampling frequencies.

Click here to book a demo for this product.

## 100x better Machine Learning CI/CD Pipeline

Lightscline Edge optimizes the entire CI/CD pipeline for model deployment, making it more efficient, faster, and cost-effective. By leveraging a small percentage of raw data for training, Lightscline Edge generates smaller models, which in turn reduces training and retraining time, and leads to significant savings in infrastructure costs.




## Conclusion

Using just 4 lines of code, Lightscline AI performs real-time data collection through prediction with >10x speed and energy efficiency over conventional approaches governed by the Shannon-Nyquist sampling theorem. This leads to orders of magnitude savings in (i) data infrastructure costs and time, and (ii) human resource efficiency. Additionally, this enables several new applications not possible today due to extreme SWaP-C requirements of real-world AI applications.

Reach out to info@lightscline.com for any queries.