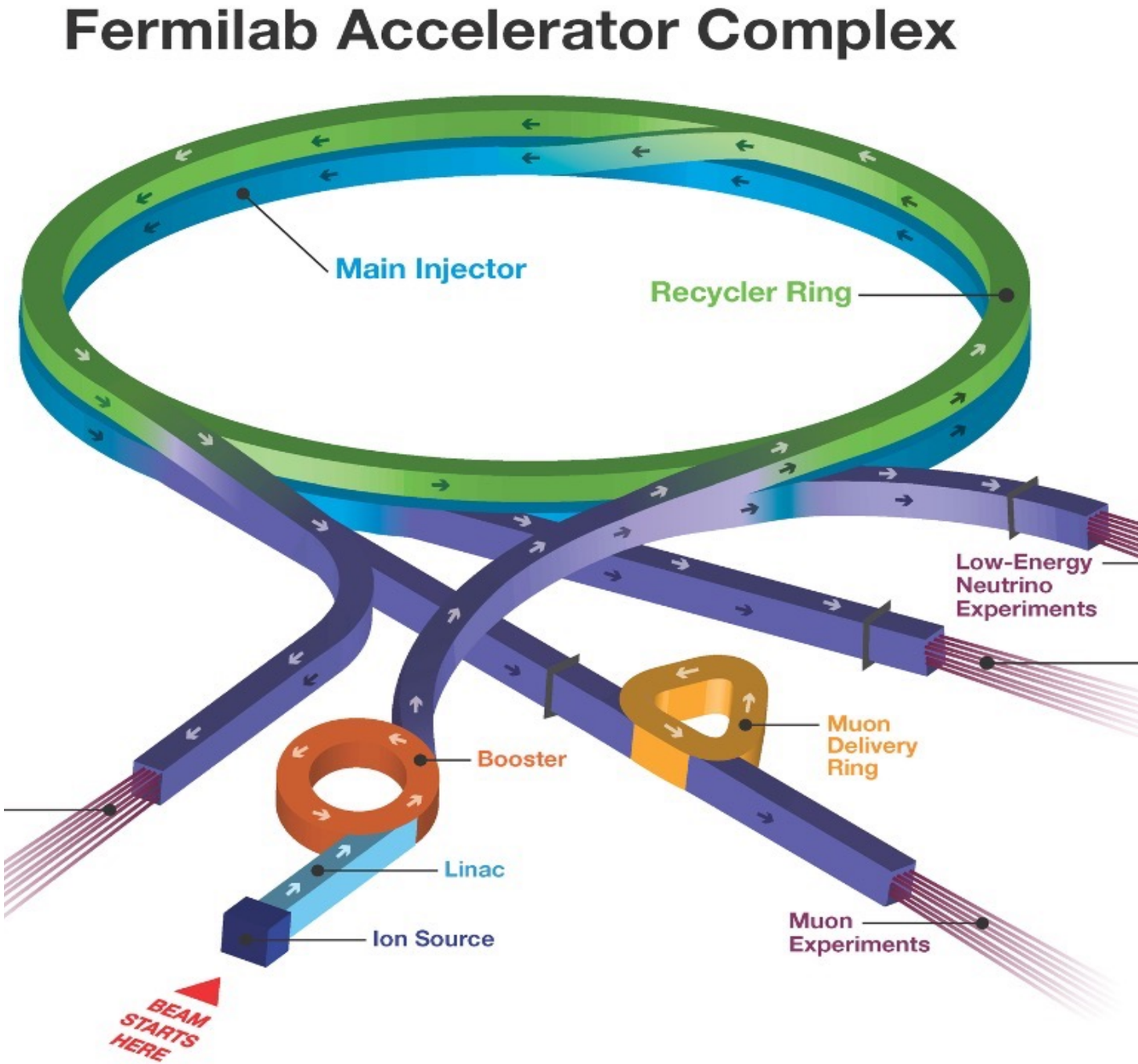
Pacific Northwest
NATIONAL LABORATORY

# L-CAPE: Linac Condition Anomaly Prediction of Emergence

November 2021

Vinay Amatya, Jamil Hasibul, Milan Jain,
Gihan Panapitiya, **Jan Strube**




U.S. DEPARTMENT OF ENERGY BATTELLE
PNNL is operated by Battelle for the U.S. Department of Energy

# PNNL Project Team

Jan Strube
Physicist

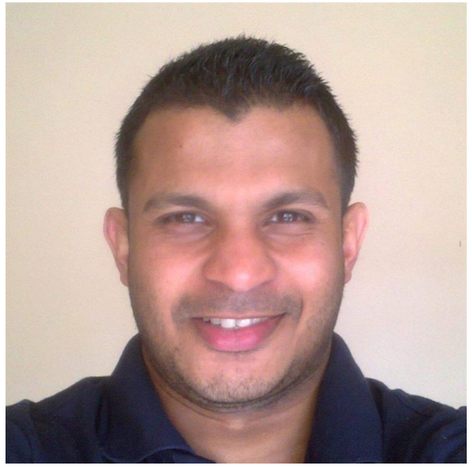
Vinay Amatya
Computer Scientist

Milan Jain
Data Scientist

+ FNAL: Bill Pellico, Jason St John, Kiyomi Seiya, Beau Harrison, Kyle Hazelwood, et al.

## Postdocs / Interns

Md Hasibul Jamil
Data Scientist

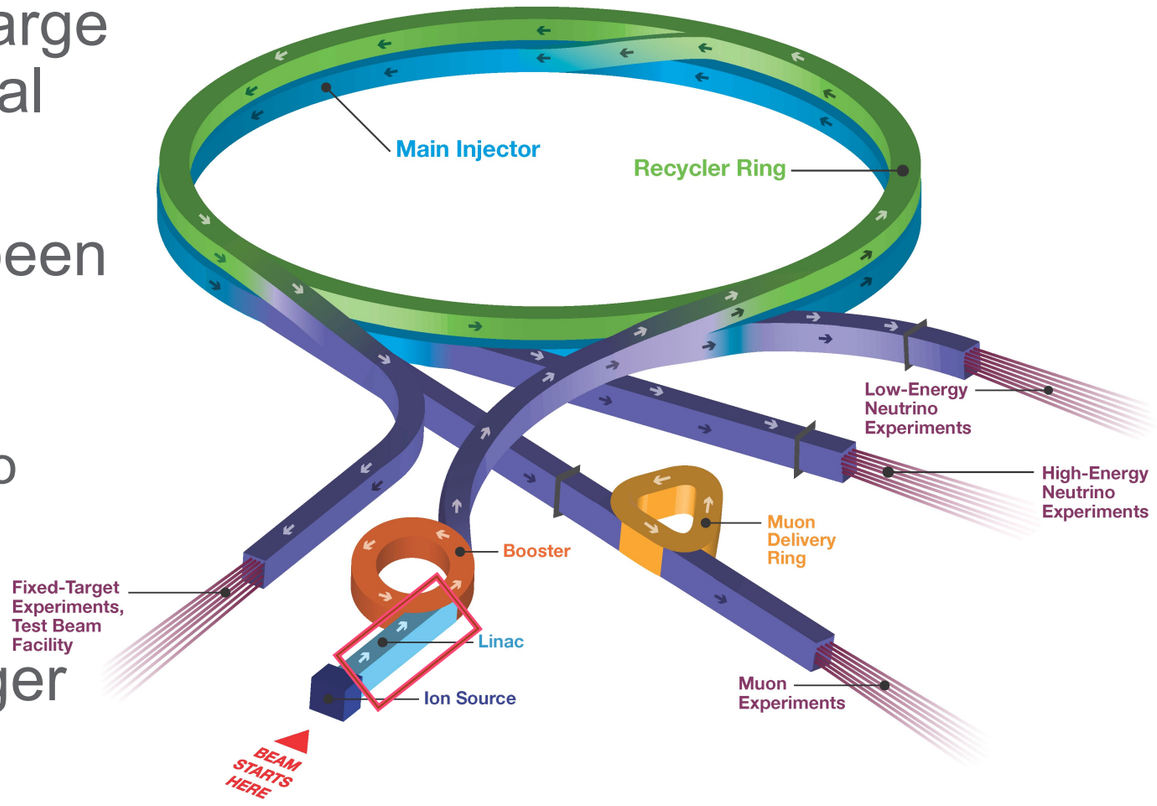


Gihan Panapitiya
Data Scientist

# Project objective

- The FNAL accelerator complex is a large scientific user facility with a substantial energy footprint
- Instrumentation of the complex has been incremental, and the controls system does not have a global view
  - Downstream systems do not respond to outages upstream
- The energy savings of downstream systems going to standby during longer outages could be quite significant

**Fermilab Accelerator Complex**



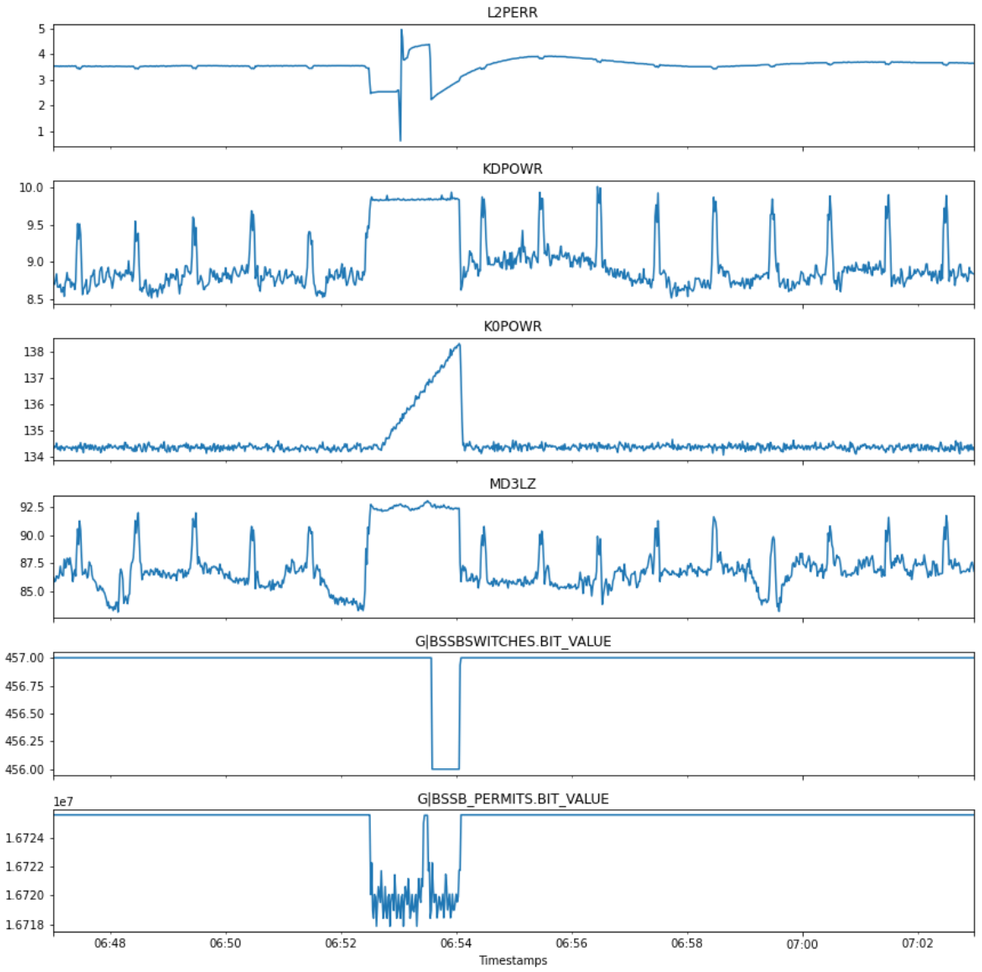
# Research Questions

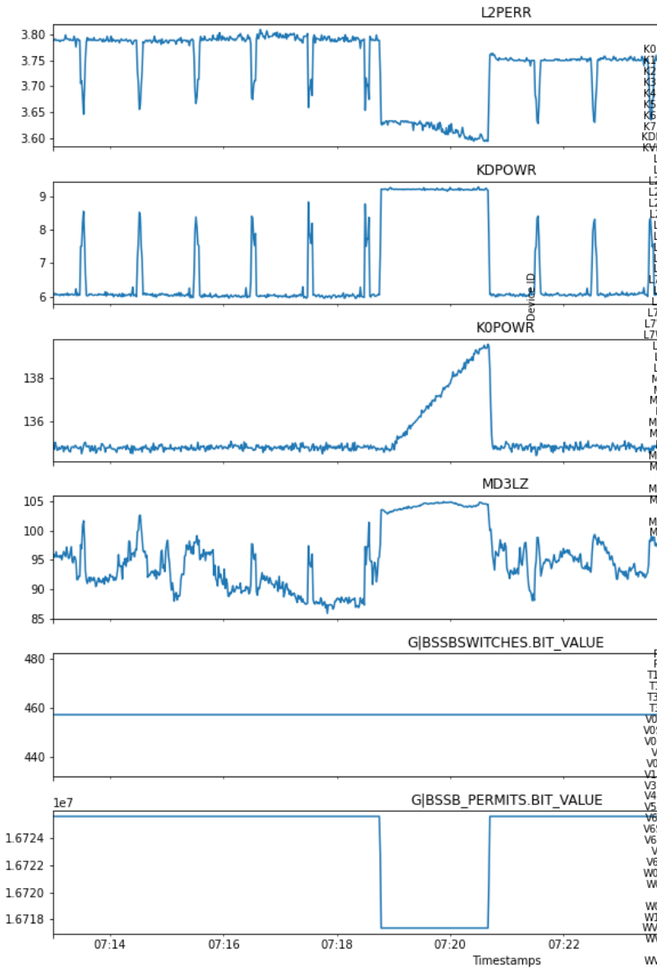
- **Can we reliably label different types of beam outages automatically?**
  - **With what accuracy?**
- **Can we reliably predict the occurrence and duration of beam outages?**
  - **How much in advance? With what accuracy?**

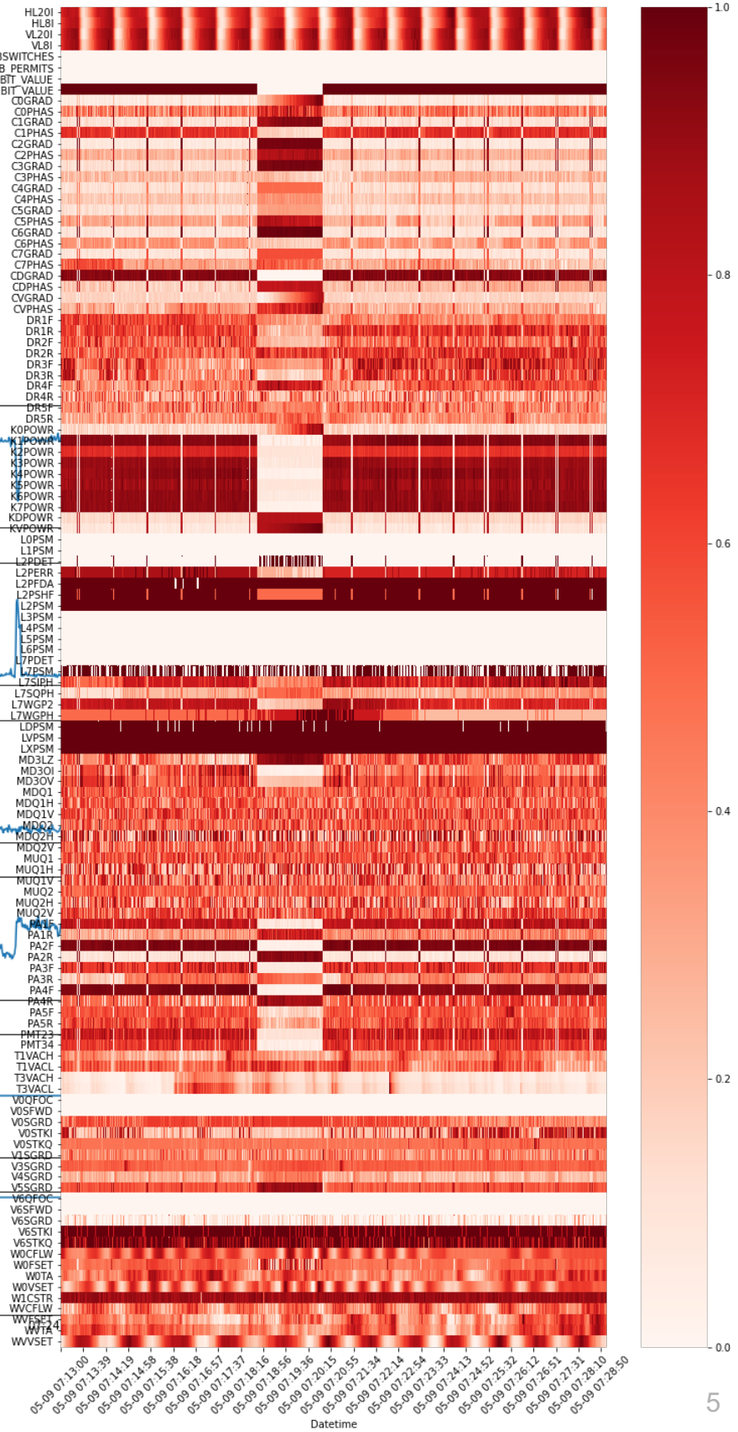
To answer these main questions, we are looking at related issues:

- How much of the data that is being stored is useful?
  - How many of the devices add no information?
  - How frequently do the devices need to be sampled?
- What kind of machine learning technique do we need to contend with short-term predictors of faults and long-term changes due to temperature/humidity ?
- How self-similar are the different systems? Can we re-use the same ML system and just feed it the same data from a different system?
  - Temperatures, voltages, currents, power, phase, …
- Are there trends / correlations between / within systems that we can exploit?

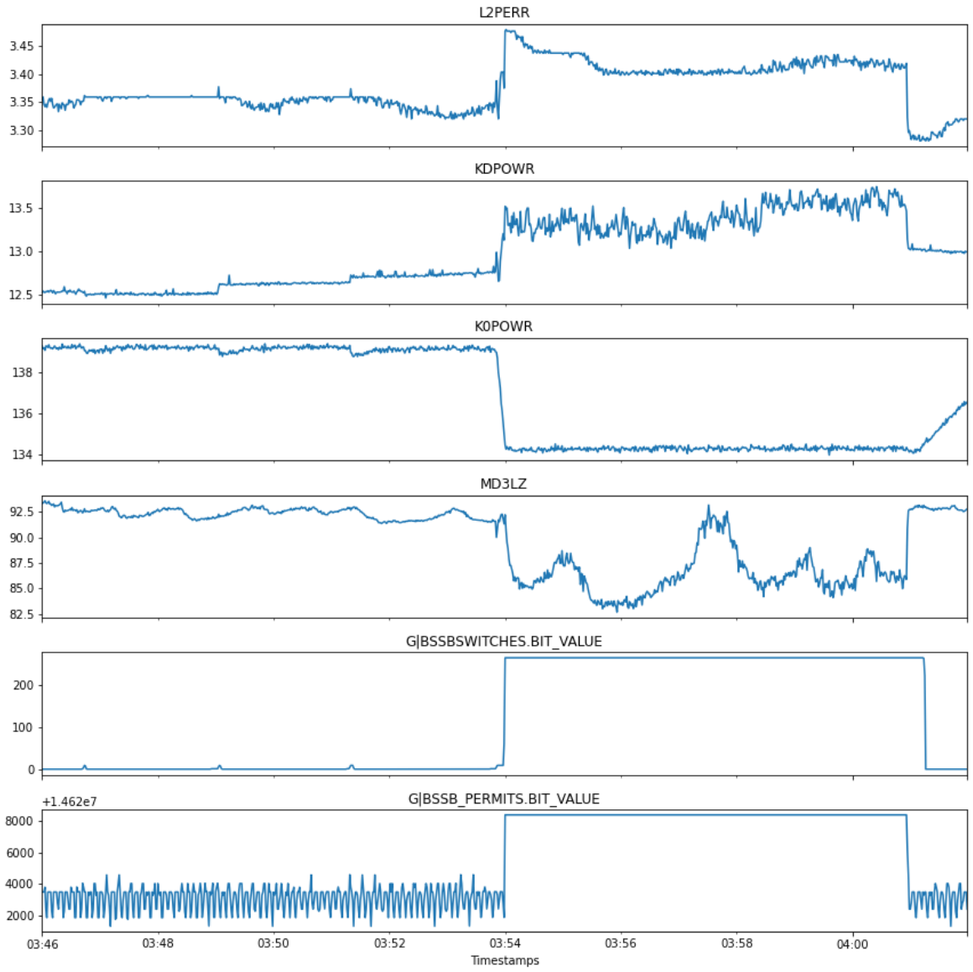
# Anomalous Patterns – KRF2 CS

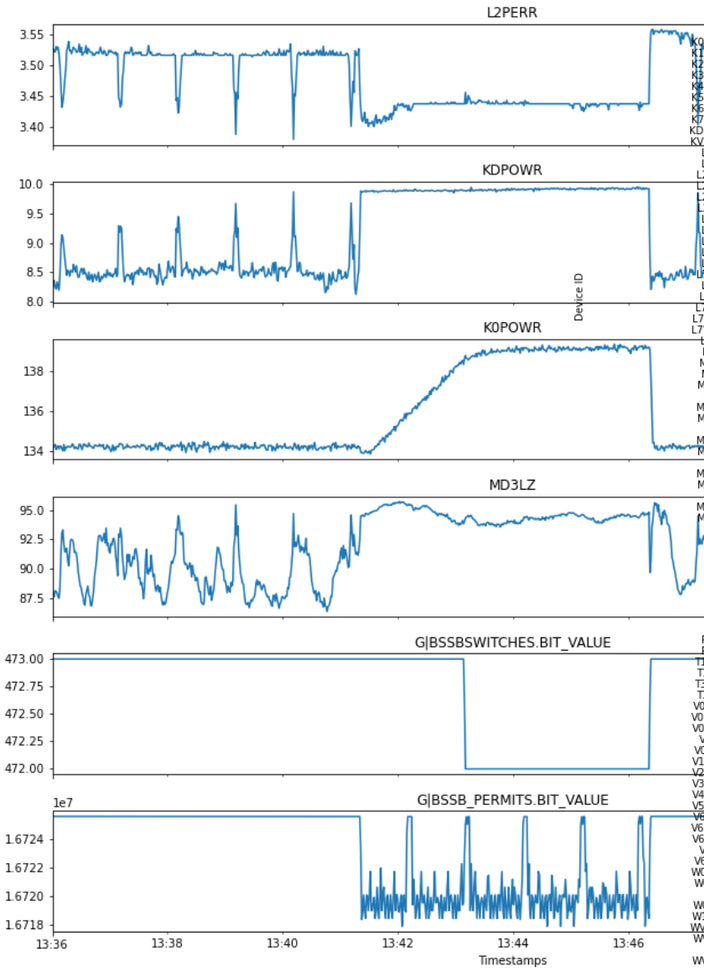


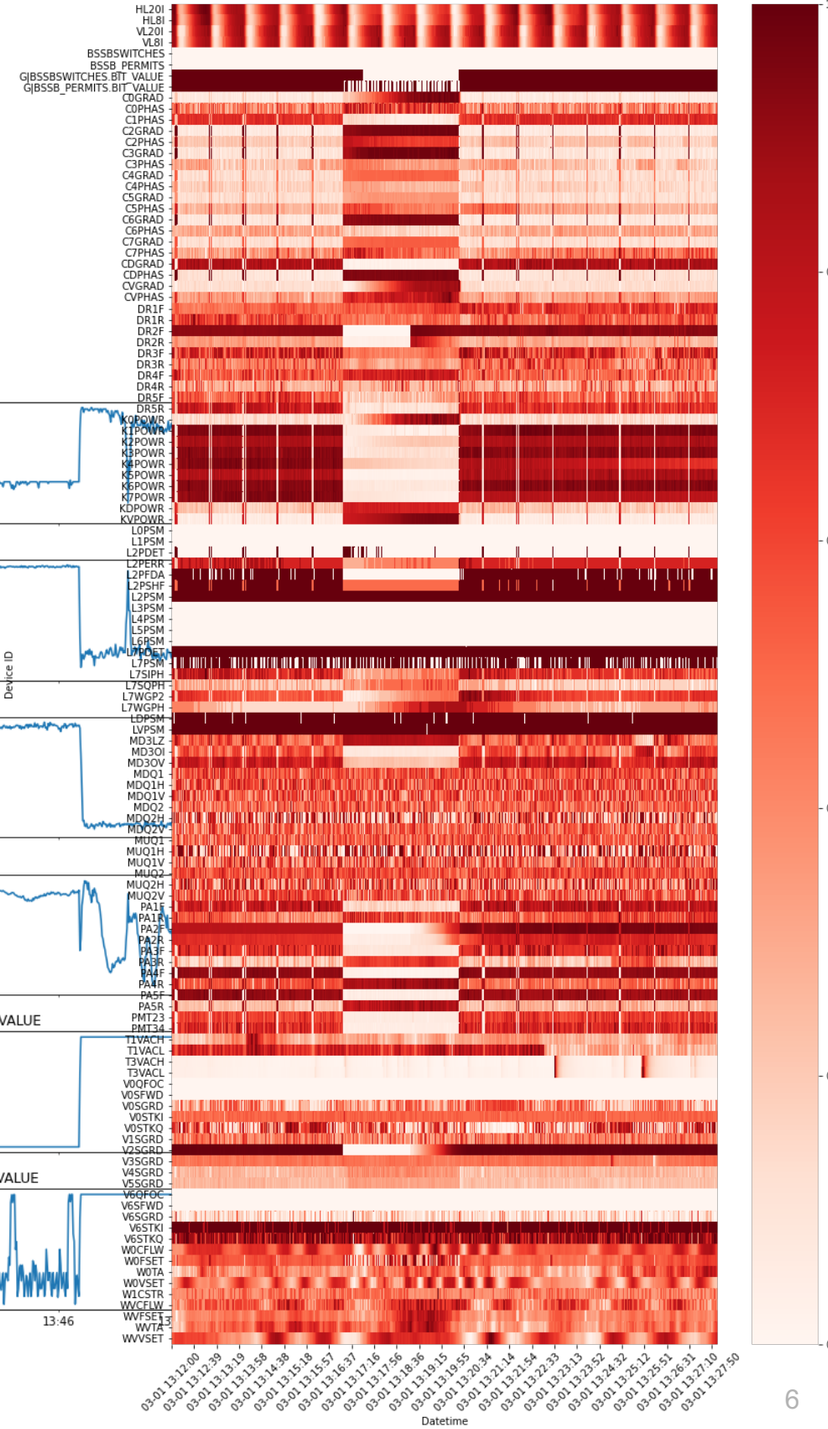



# Anomalous Patterns – LRF2 Overvoltage



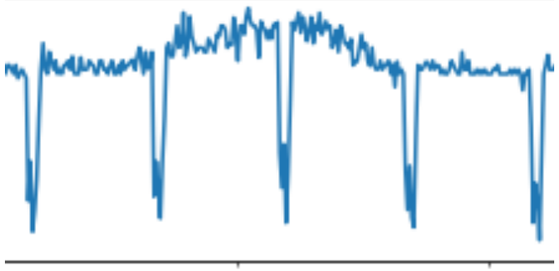
# Real world data sets are messy

- 2945 different columns (devices), 2208 unique
  - Some devices are sampled at different frequencies
- Devices are sampled at different frequencies:
  - 98 devices sampled at 66 ms
  - 567 devices sampled at 1 s
  - 118 devices sampled at 10 s
  - 763 devices sampled at 15 s
  - 316 devices sampled at 30 s
  - 746 devices sampled at 120 s
- Time steps are not aligned
- There are gaps in the data (sometimes big gaps, > 10k time steps)
- Labels are subject to human error (typos, omissions, …)
- ~600 MB / h x 7600 h; 3.6 TB total

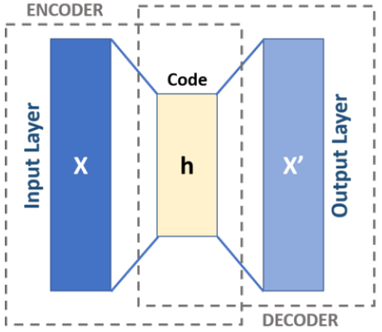
# Data analytics for the FNAL complex – overview

Pacific Northwest NATIONAL LABORATORY

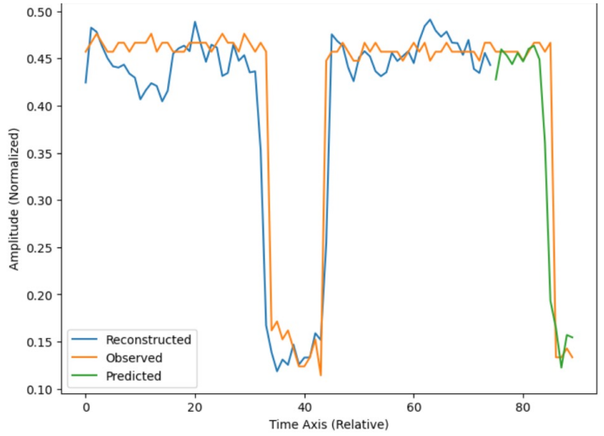
## Time-series data from devices

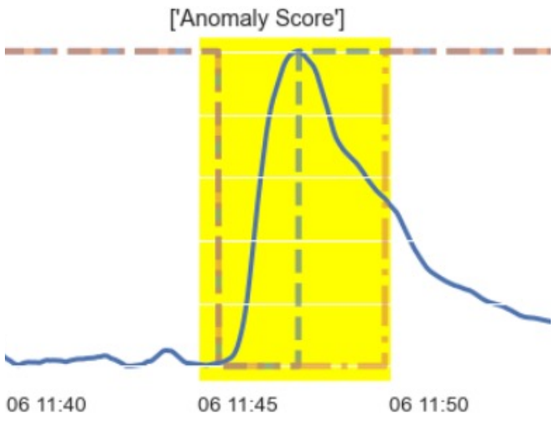


## LSTM autoencoder

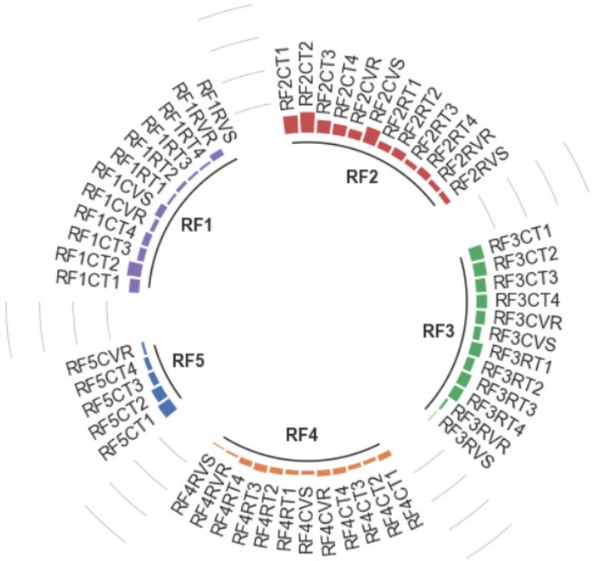



## Reconstruction error

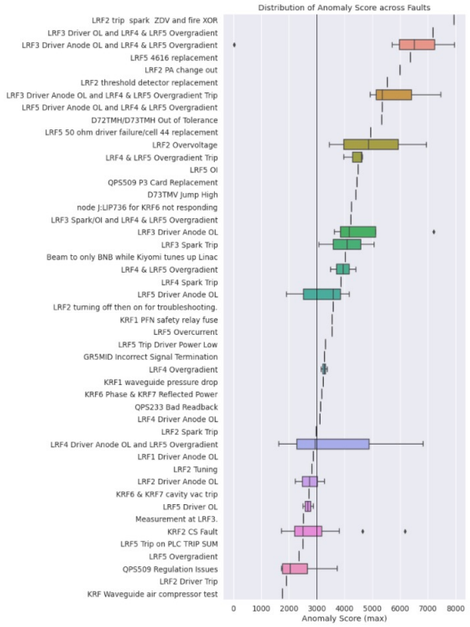



Combined anomaly score from all devices

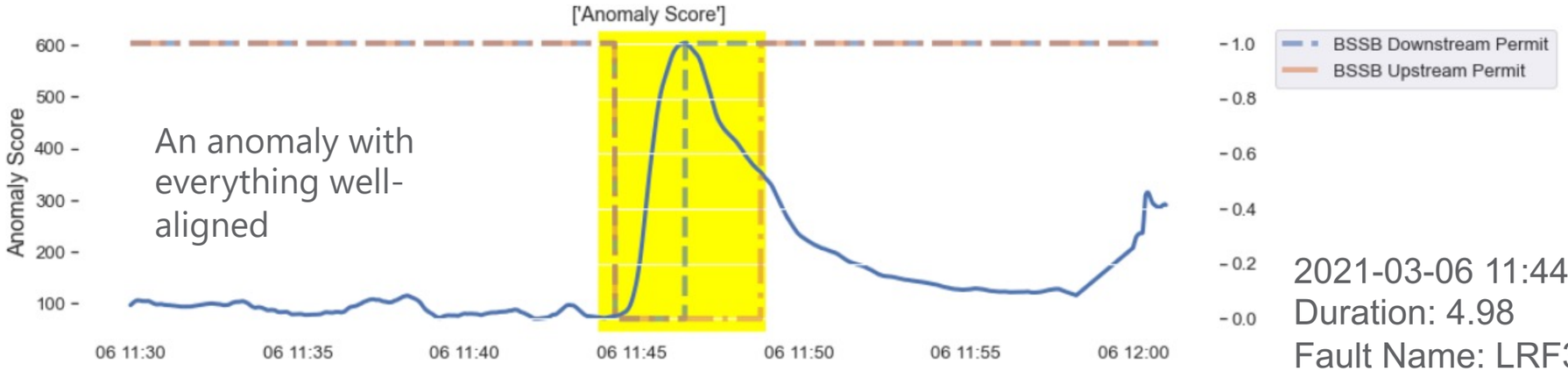



Breakdown of devices contributing to a given anomaly



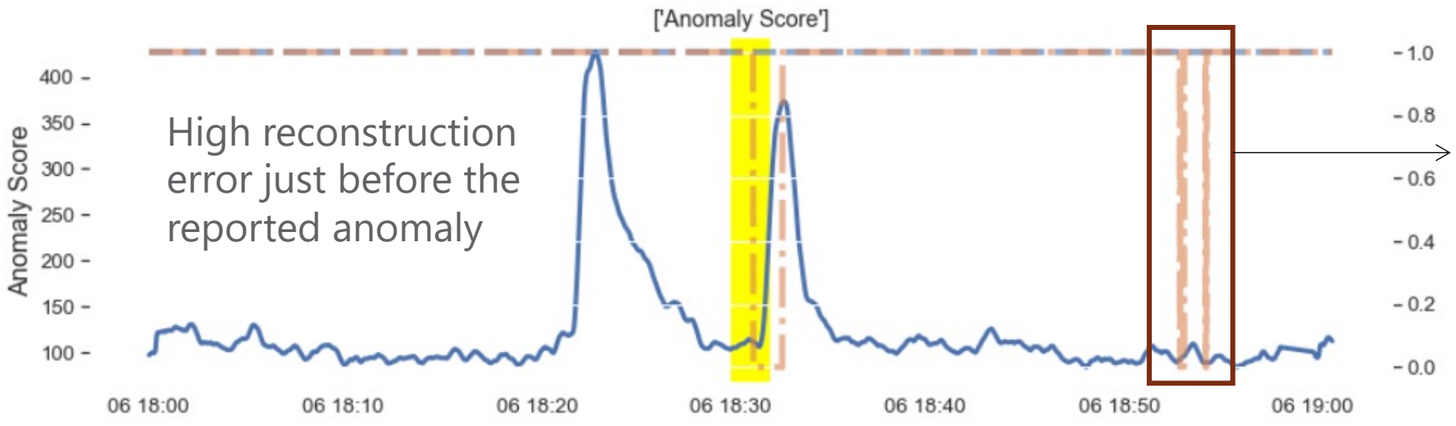

Thresholding to catch anomalies

# Detection using Anomaly Score




An anomaly with everything well-aligned

2021-03-06 11:44:00
Duration: 4.98
Fault Name: LRF3 spark trip



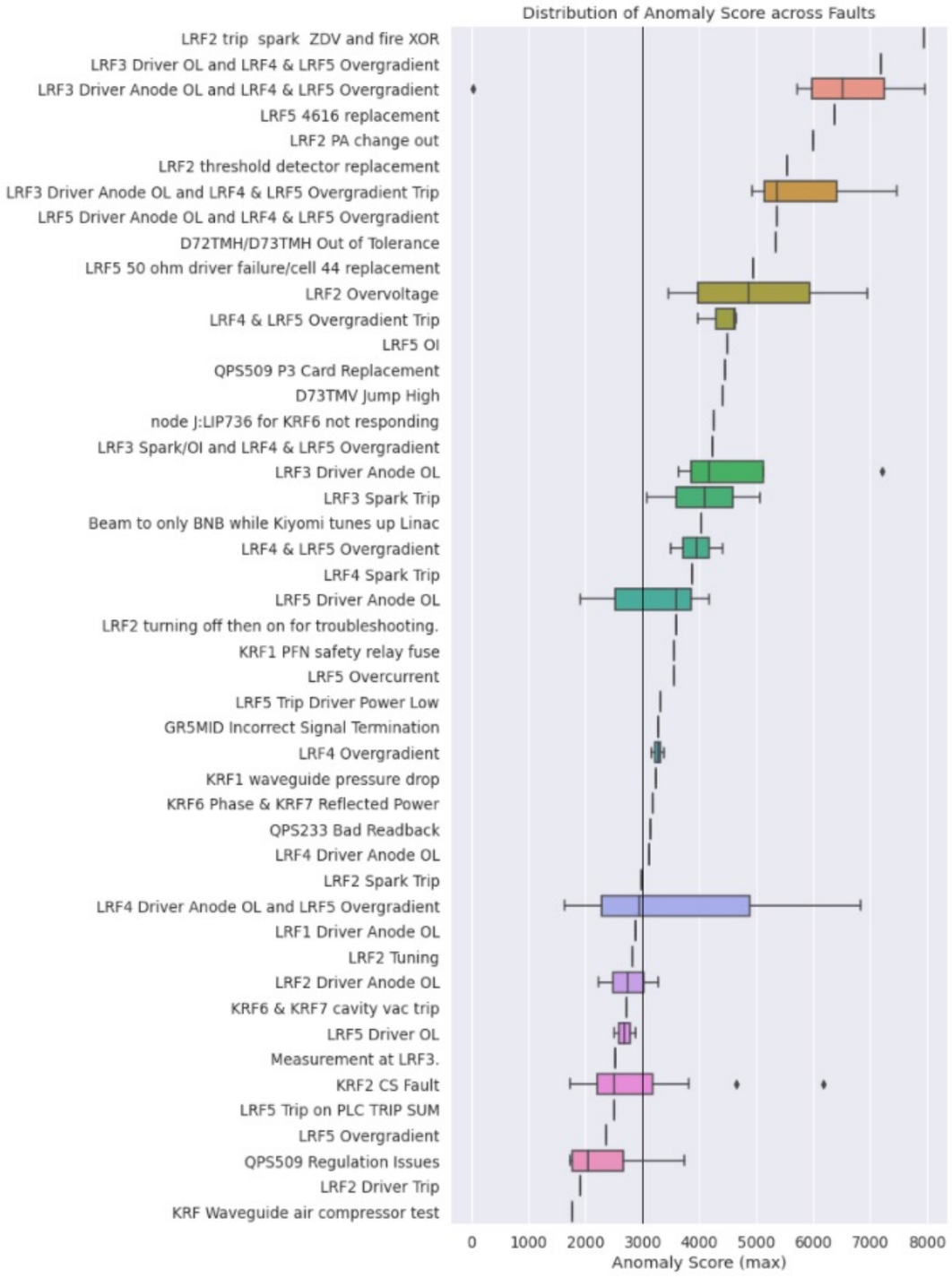

High reconstruction error just before the reported anomaly

Bits went down but reconstruction error was fine, and no fault was reported

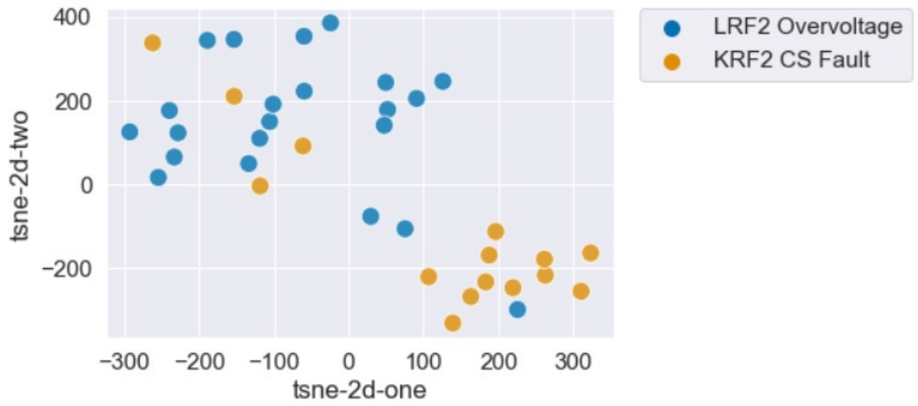
2021-03-06 18:30:00
Duration: 1.98
Fault Name: LRF3 spark trip

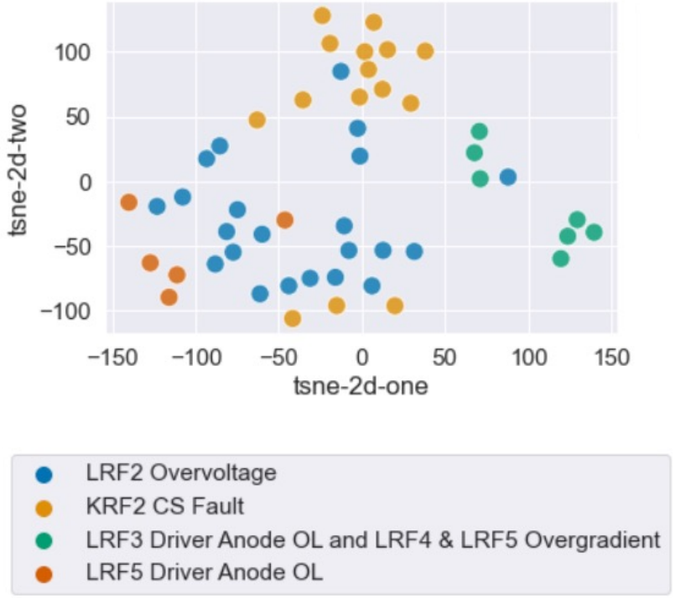
# Identification of Faults

- The plot shows the different labels that operators have assigned the outages
- The x-axis is the anomaly score that we attribute to each fault
  - Not all faults occur with the same frequency – we show the distribution of anomaly scores for each type of outage
- To identify faults automatically, we apply a threshold



# Unsupervised Fault Labeling

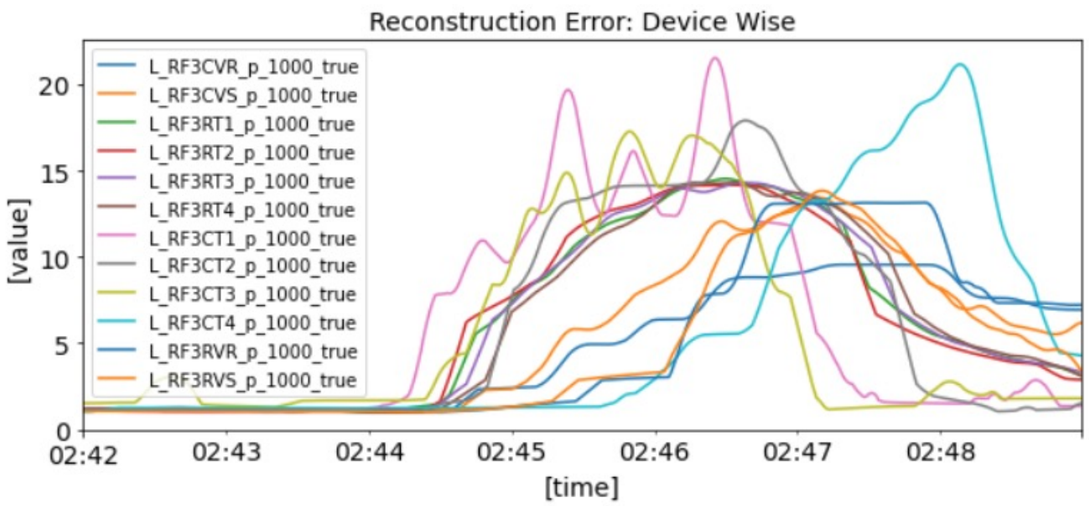




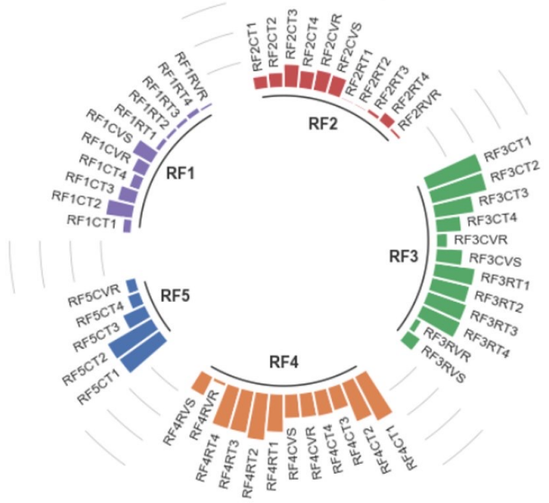

First look at our ability to distinguish between faults, using t-SNE

- Few, frequently occurring outages can be clustered by eye
- Adding more classes of outages with fewer occurrences complicates things
- Improvements will come from
  - Better clustering: Self-Organizing Maps; UMAP Clustering
  - Adding the time component of the outages

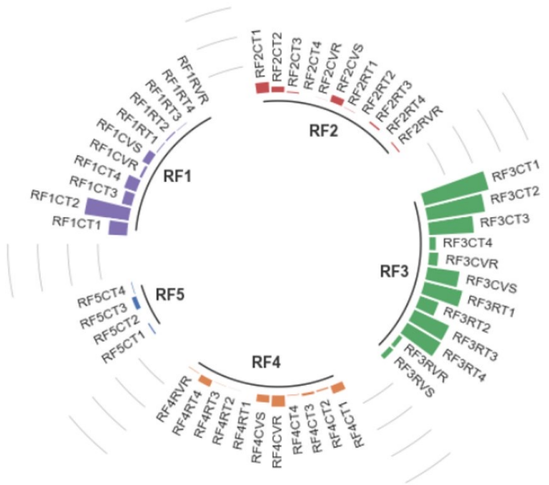
# Avenues for further classification improvement: Time dependence and fingerprinting




LRF3 Driver Anode OL and LRF4/5 Overgradient




LRF3 Driver Anode OL



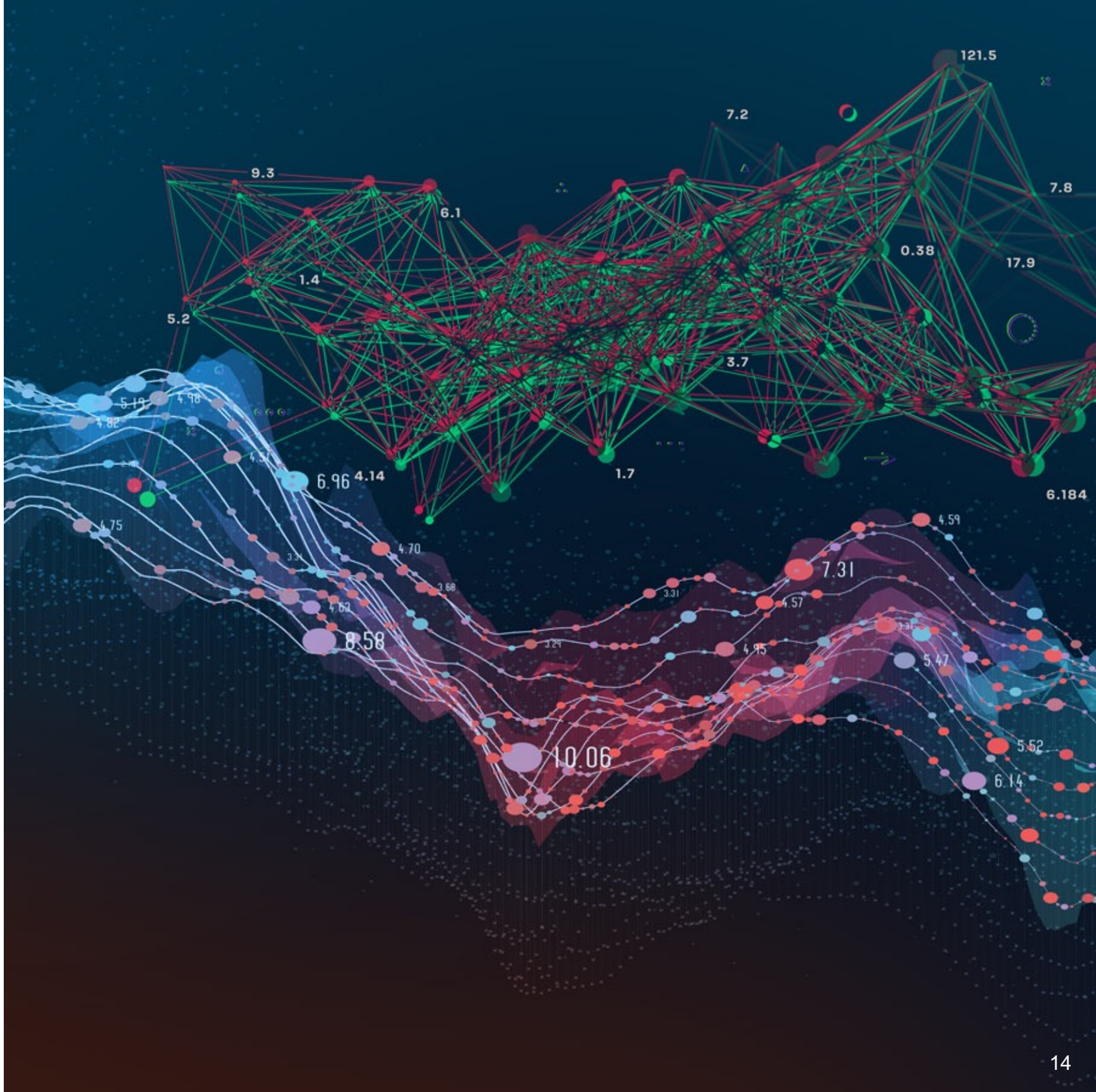

Outages have a time component, where different devices respond with a different time constant

Different types of outages have different responses in different subsystems. We are working on exploiting this physical structure to improve labeling.

# Status and Next Steps

- Data cleaning and preparation routines are defined.
- The whole analysis chain prep → model → outage identification exists.
- We see some examples of the clustering working.
- There are still too many false positives. Improvements under study include
  - Better de-noising / modeling of devices
  - Pruning of devices in the anomaly score
  - Exploiting the time structure of outages.

- We will start to look at the forecasting. The current anomaly score is not powerful enough to predict outages and only starts deviating significantly after the start of the outage.

Pacific Northwest
NATIONAL LABORATORY

# Thank you