Application of machine learning to find anomalous events in LZ data

Chami Amarasinghe - 05/12/2022
CoSSURF 2022

Work with Scott Kravitz, Maris Arthurs, and Yi Liu
On behalf of the LZ Collaboration

UNIVERSITY OF MICHIGAN

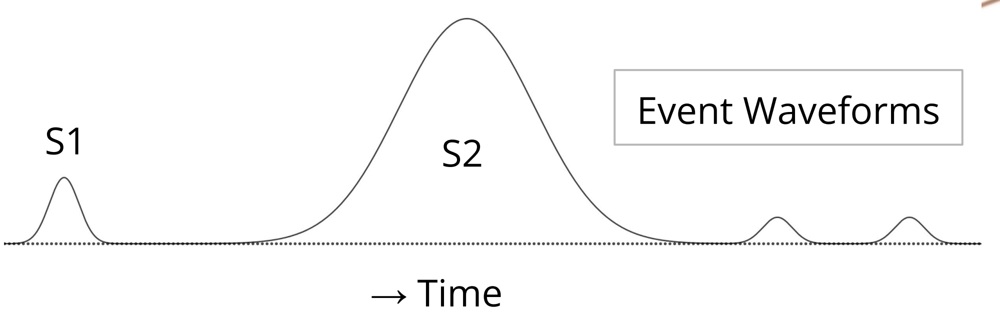
The LZ Dark Matter Experiment

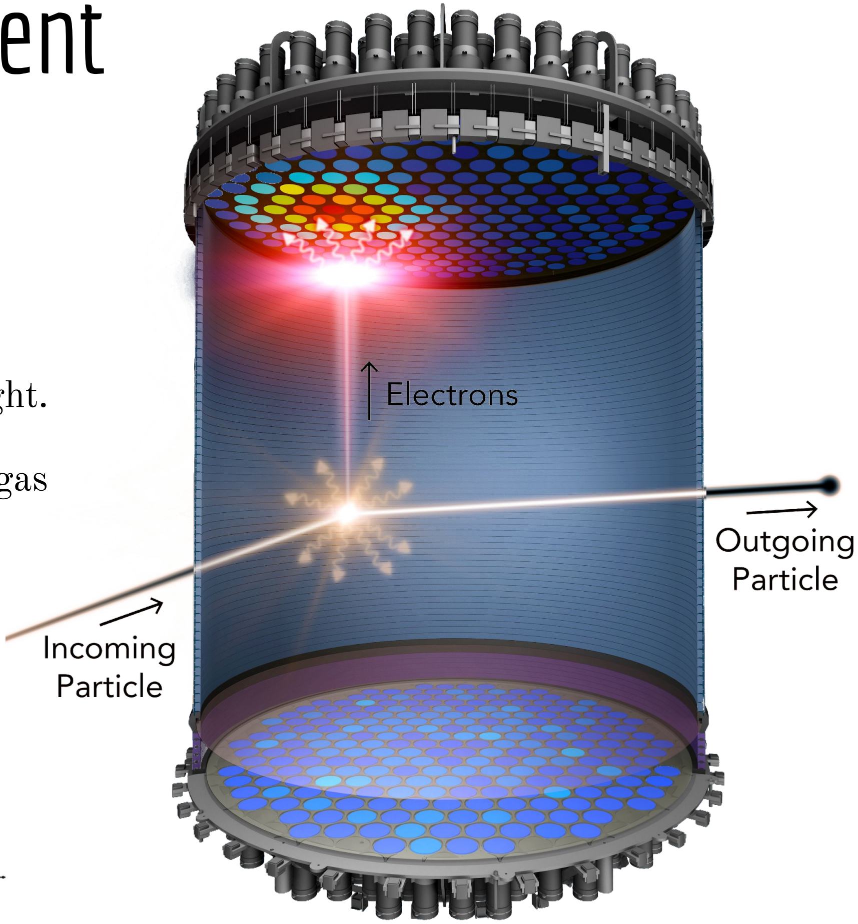
LUX-ZEPLIN (LZ) is an underground direct detection experiment at SURF.

Particle interactions with liquid xenon produce two signals:

S1 - Scintillation - Initial interaction causes LXe to emit light.

S2 - Ionization - Electrons are drifted and extracted into a gas Xe layer, which scintillates.





Anomaly Finding in LZ

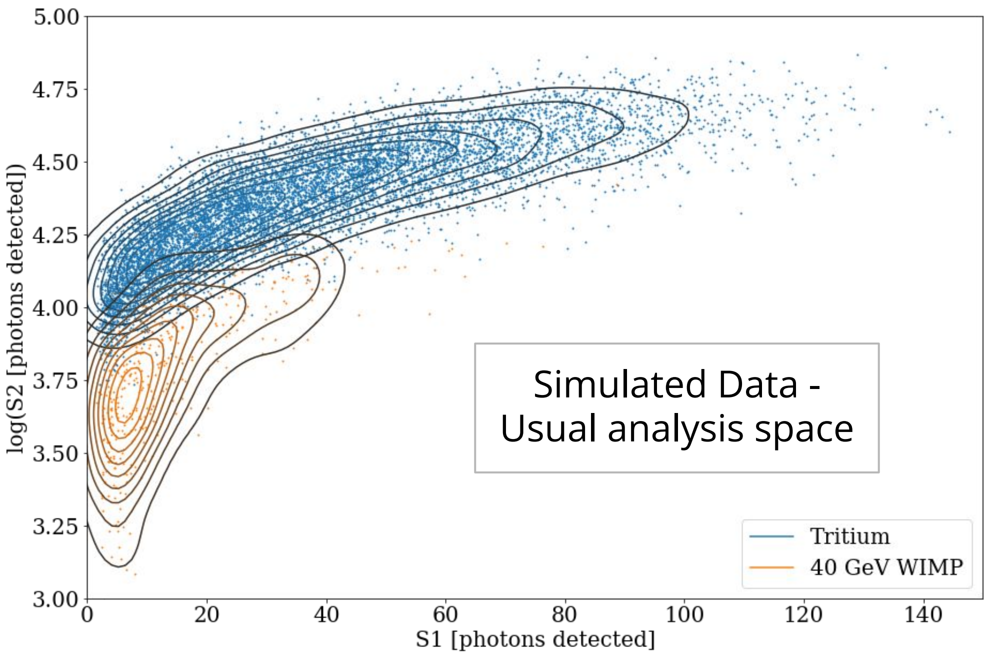
Goal - Quickly identify and interpret anomalous data in high-dimensional spaces.

Features

- Pulse shape and size.
- 3D position.
- Signal distributions.
- Number of pulses in event.

Use Cases

- Rare background discrimination.
- Tuning aid for simulations and data processing algorithms.
- Waveform handscanning aid.
- Detector anomalies in real data.



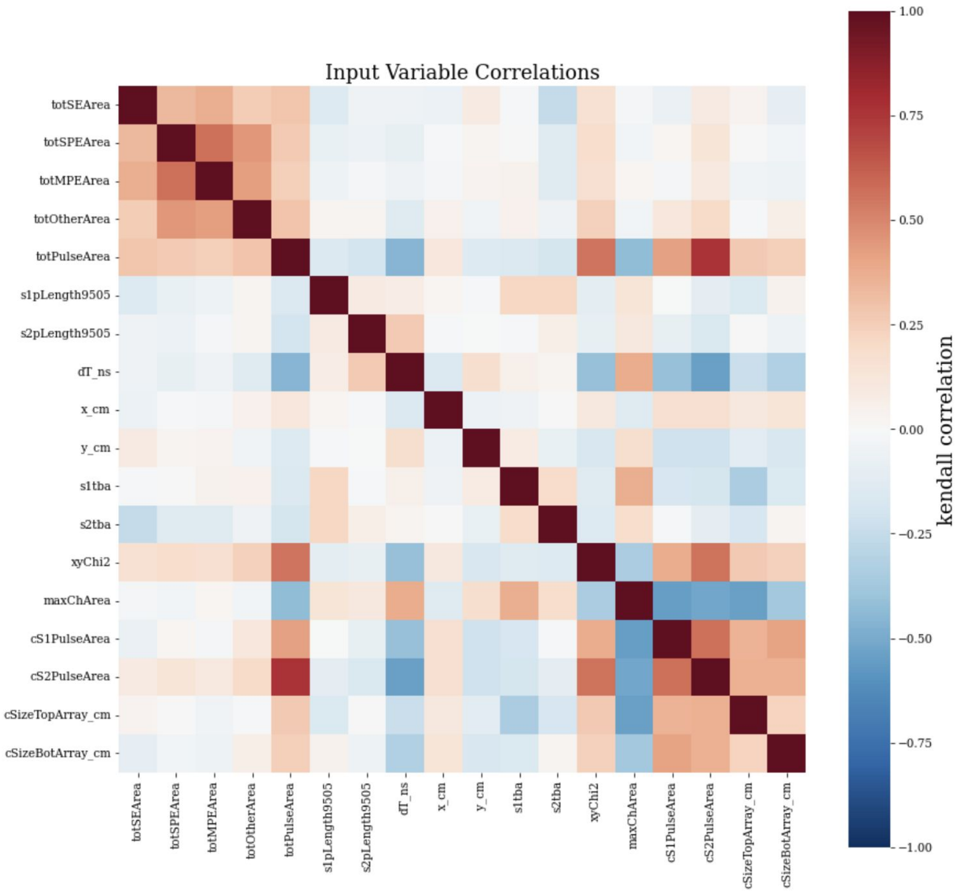
Taking advantage of the high-dimensional feature space, we have explored two unsupervised learning techniques for finding patterns in the LZ data.

1. Isolation forest
2. Dimensional reduction & clustering

LZ Data Space

Data contains both pulse and event information

- **Event level features**
  - Total area of different pulses in the event
    - Single electrons, single photoelectrons, etc.
- **S1 & S2 pulse features**
  - Pulse length
  - Pulse area
  - Summary of pulse shape
- **Other features**
  - S1 & S2 top bottom asymmetry (TBA)
  - Drift time
  - XY position
  - S1 hit pattern size



1. Isolation Forest

The isolation forest is an ensemble of random decision trees.

1. Starting at the root node, a uniformly random cut is applied to a random feature.
2. Repeated recursively to build a tree, until the datum is isolated from others.
3. Outliers take fewer cuts to isolate.

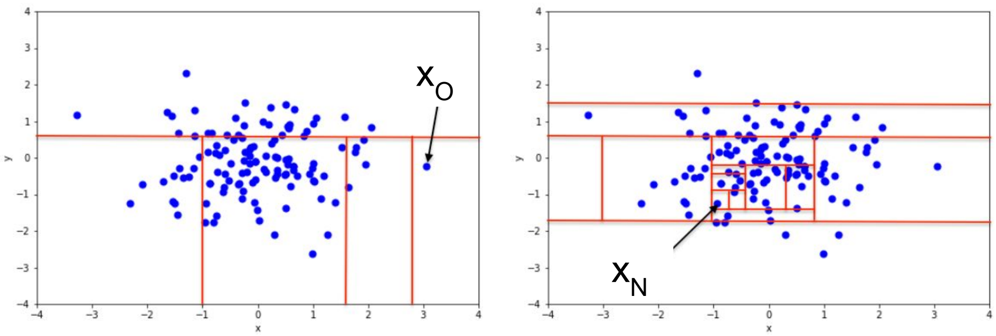
Anomaly Score - Function of the length of decision path.

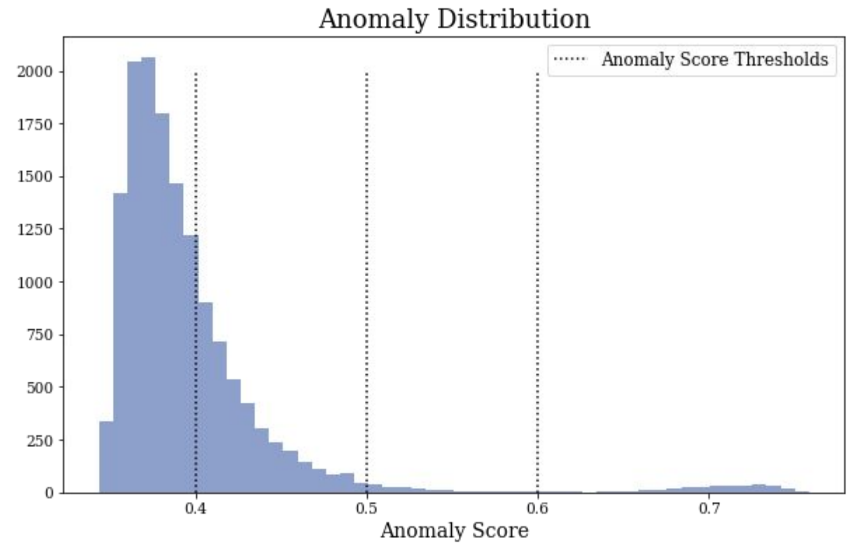
Why is a certain event anomalous?
Why is a certain set of events anomalous?

This technique is directly interpretable - Features that are cut on frequently are the cause of the outlier.

FT Liu, et. al., *Isolation forest*, ICDM 2008.

2D Example





1. Isolation Forest

The isolation forest is an ensemble of random decision trees.

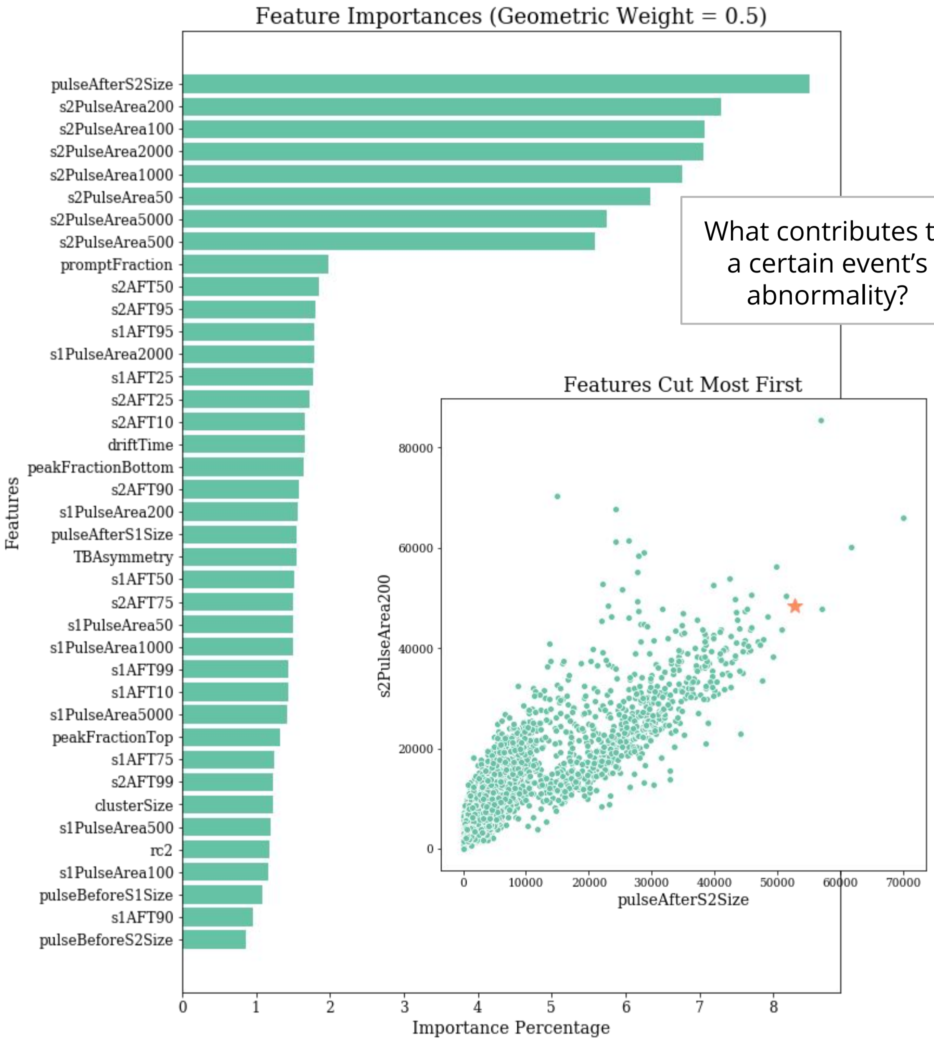
1. Starting at the root node, a uniformly random cut is applied to a random feature.
2. Repeated recursively to build a tree, until the datum is isolated from others.
3. Outliers take fewer cuts to isolate.

Anomaly Score - Function of the length of decision path.

Why is a certain event anomalous?
Why is a certain set of events anomalous?

This technique is directly interpretable - Features that are cut on frequently are the cause of the outlier.

FT Liu, et. al., *Isolation forest*, ICDM 2008.



What contributes to a certain event's abnormality?

2. Dimensional Reduction

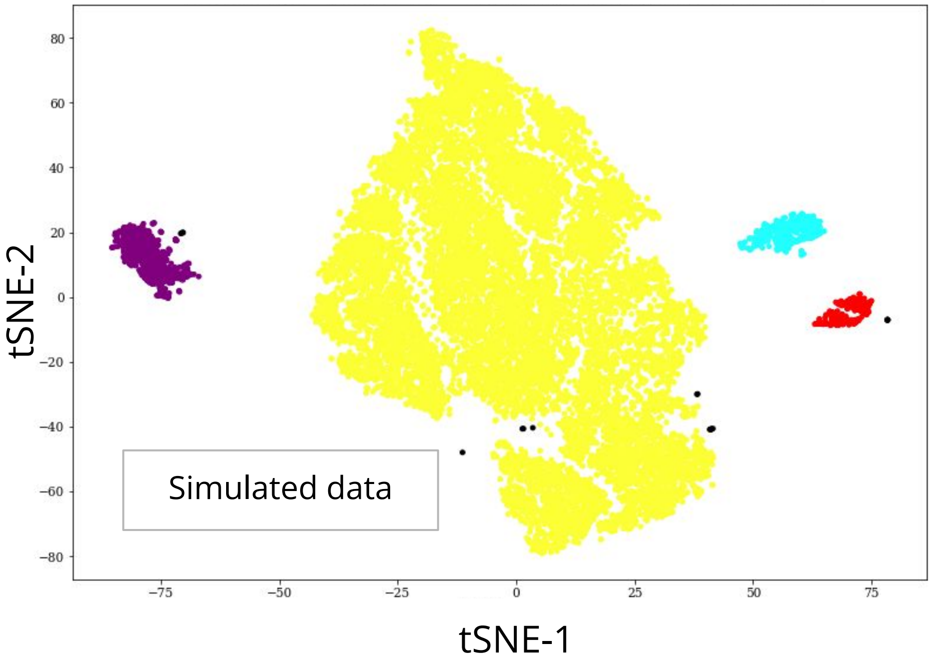
- Map N-dimensional (~30 features) data to 2D representation
    - **Why** – Outliers in multidimensional feature spaces are difficult to detect visually.
    - **Goal** – quickly identify and study (not remove) outlier events.
    - **How** – represent in 2D while **preserving structure**.

- Linear techniques preserve global structure, but lose information about local structure.
    - Example: Principal Component Analysis (PCA)
- Non-linear techniques tend to **preserve** local structure as well as global structure
    - **t-SNE**: T-distributed Stochastic Neighbor Embedding
        - L.J.P. van der Maaten and G.E. Hinton. *Visualizing High-Dimensional Data Using t-SNE*. Journal of Machine Learning Research 2579-2605, 2008.
    - **UMAP**: Uniform Manifold Approximation and Projection.
        - L McInnes, J Healy. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, ArXiv e-prints 1802.03426, 2018

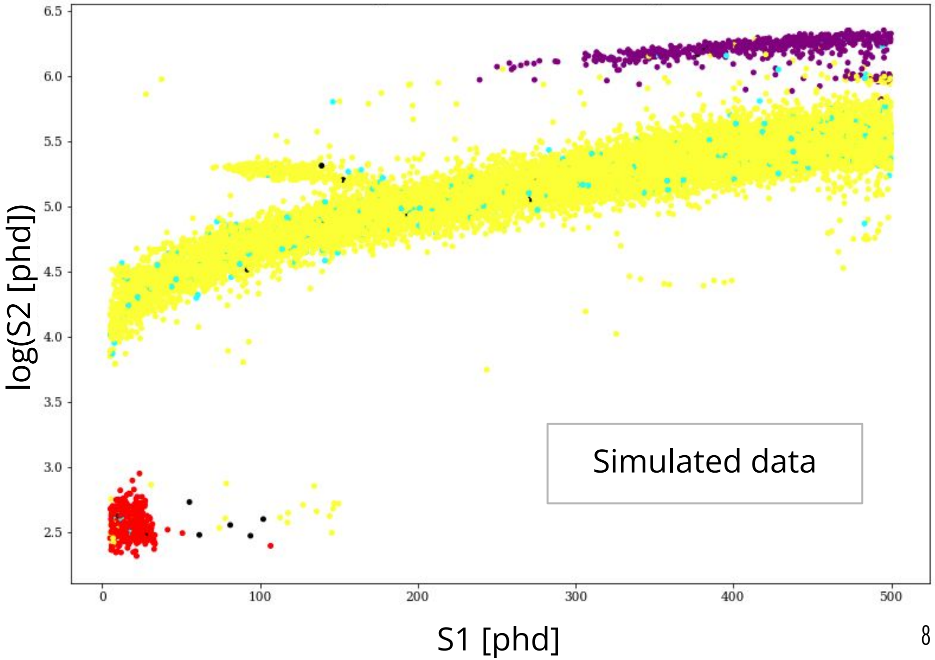
2. DR & Clustering - Simulated Data

Goal - Visualize ~30 dimensional feature space in 2D clusters and discern reasons for clusters.

tSNE-reduced data



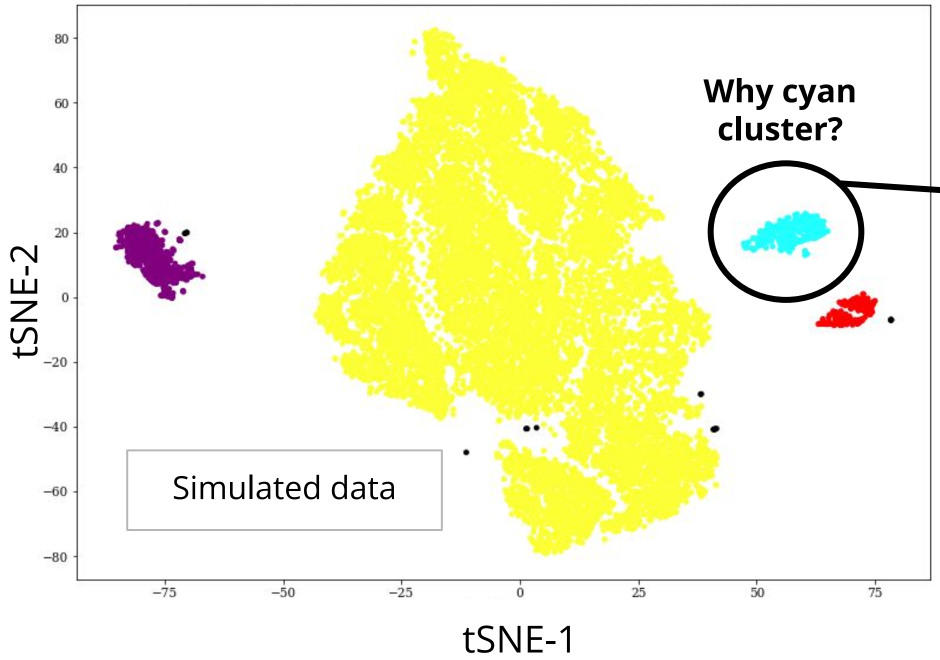
tSNE clusters in signal space



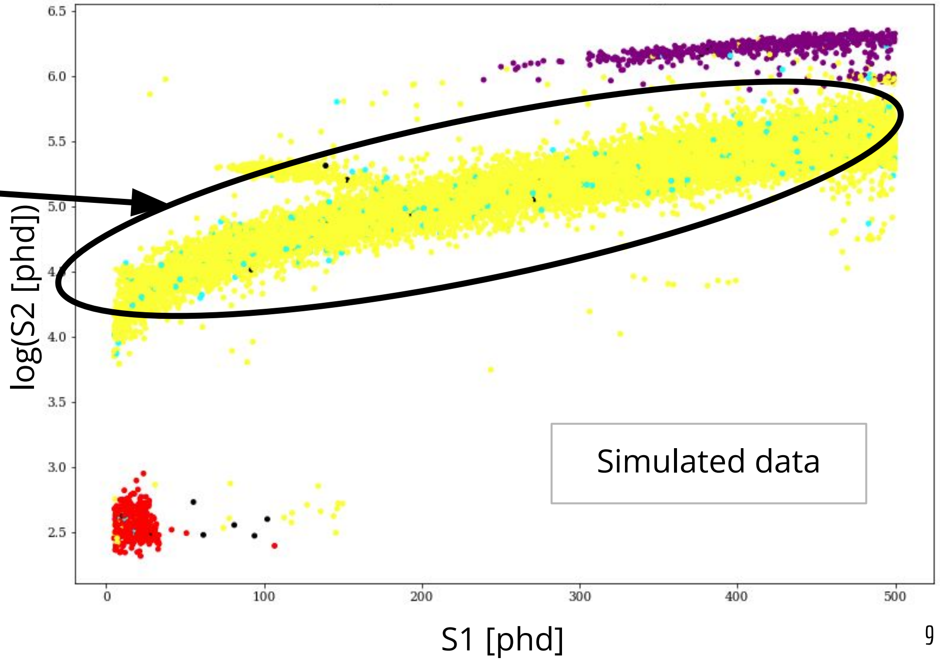
2. DR & Clustering - Simulated Data

Goal - Visualize ~30 dimensional feature space in 2D clusters and discern reasons for clusters.

tSNE-reduced data



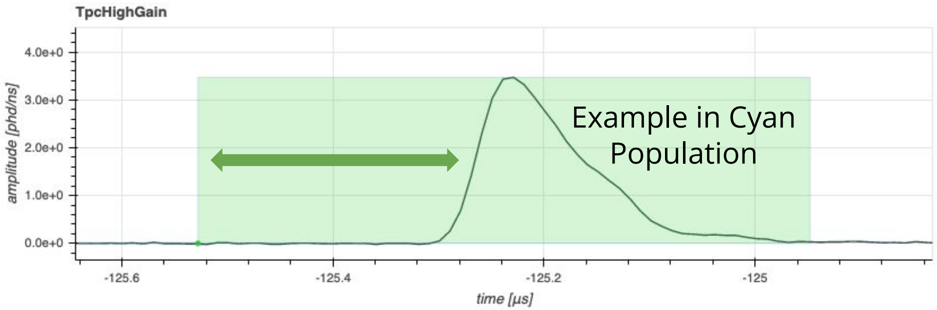
tSNE clusters in signal space

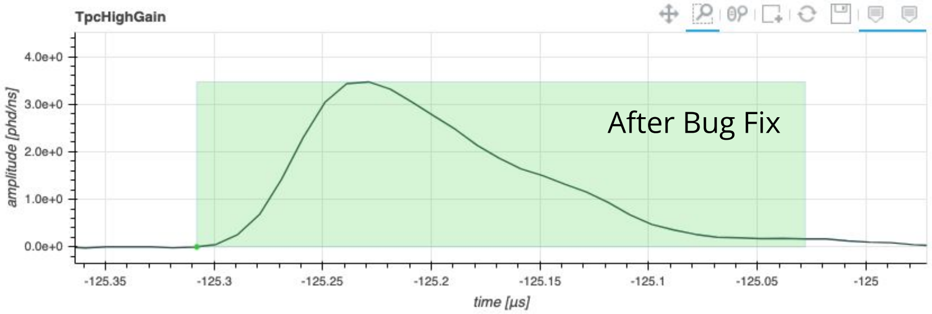


Cases in simulated data

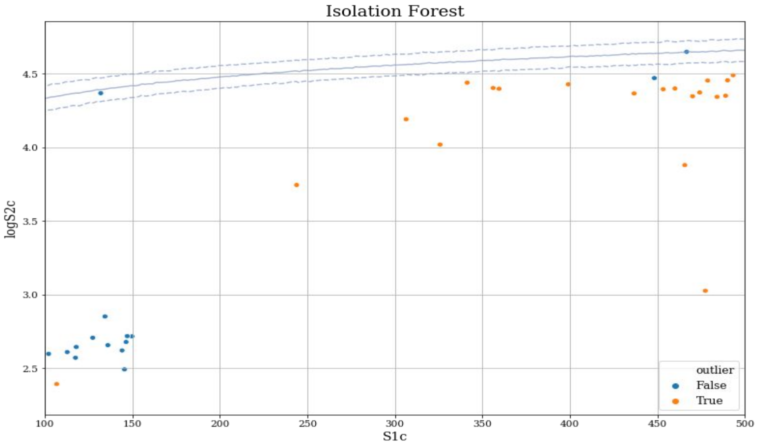
Pulse Finder Inefficiency

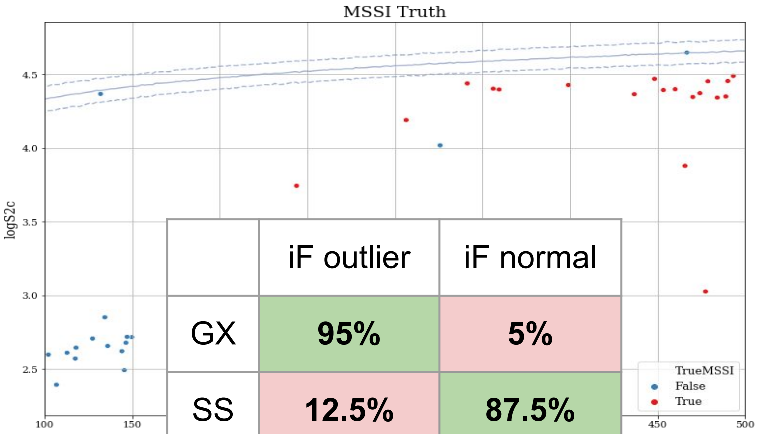
The cyan population in simulations consisted of pulses that were tagged with a long rise time.





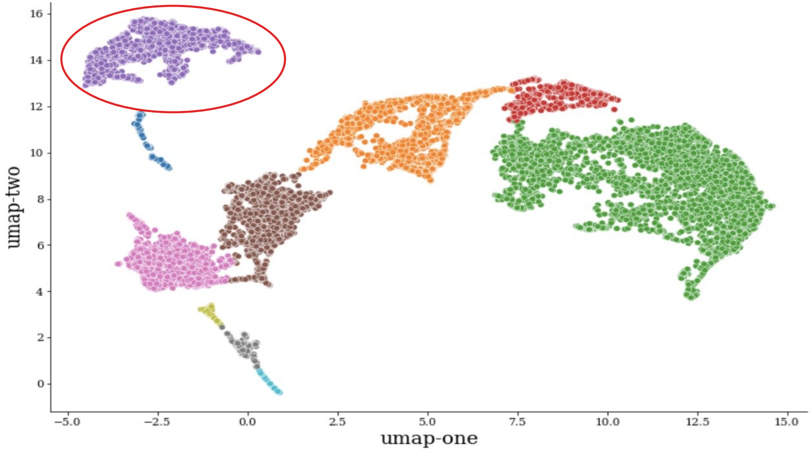
Rare Background Identification

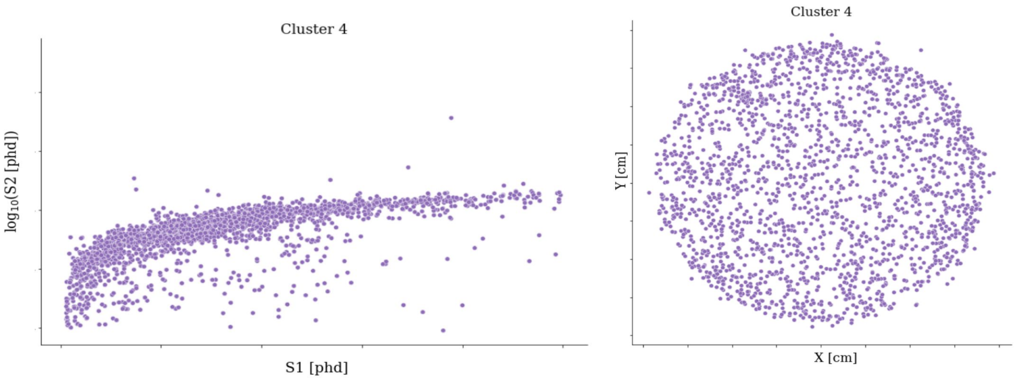


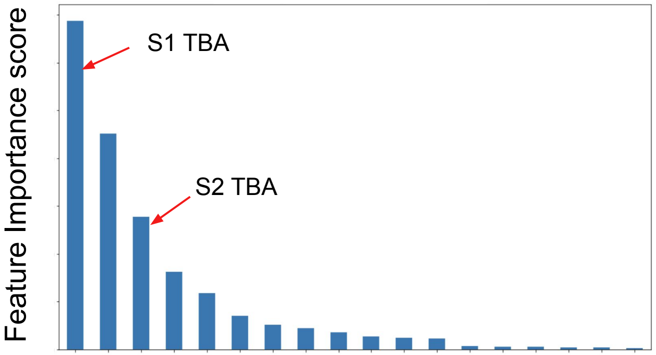


	iF outlier	iF normal
GX	**95%**	**5%**
SS	**12.5%**	**87.5%**

Clusters in Real Data - Gas Events



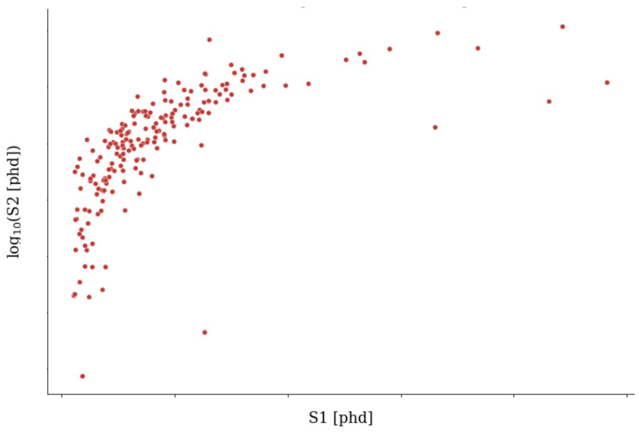


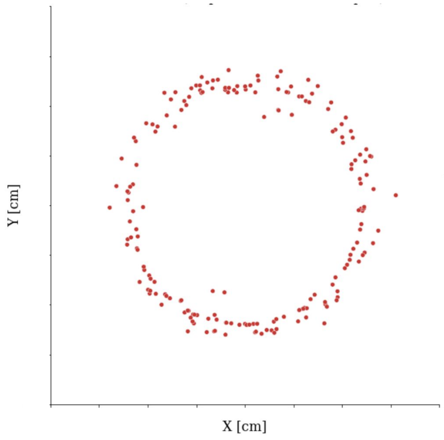


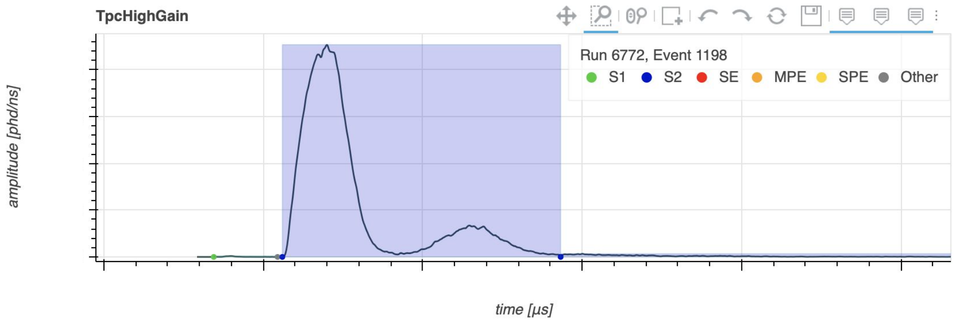
Purple cluster found to be gas events

- Found to have large S2 TBA and large S1 TBA.
- Importances allows identification of relevant RQs.

Clusters in Real Data - Photoionization







New population → after adding S2 pulse shape features

- Photoionization of the wire electrodes from S2 light
- Originates near the walls in extraction region
- Detector effect

Conclusions & Outlook

Unsupervised techniques have the potential to probe the **known unknowns** and the **unknown unknowns** in science data.

- Unknown unknowns - Use the largest representative feature set available.
- Known unknowns - Use appropriate features for the task.

Interpretability is important for studying events or groups of events. These techniques allow for a better understanding of the data.

Applications include

- Data quality,
- Anomalous backgrounds,
- Tuning data processing algorithms,
- Fixing simulation bugs.

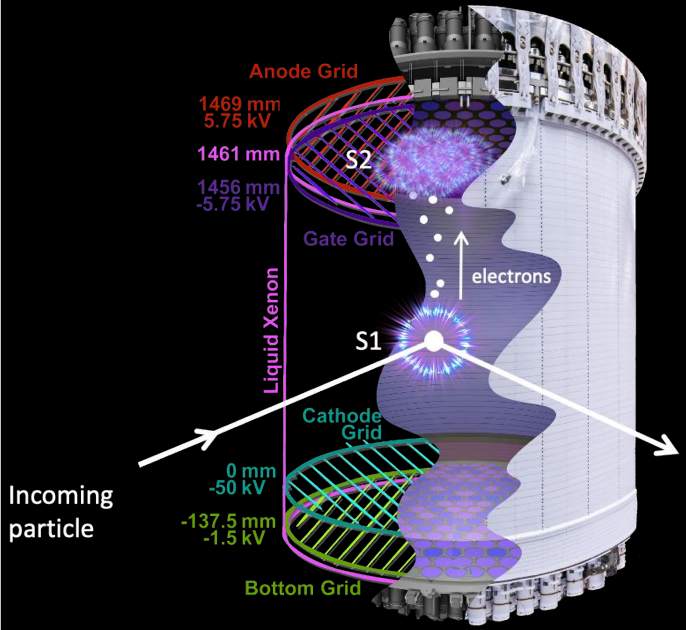
LZ (LUX-ZEPLIN) Collaboration

35 Institutions: 250 scientists, engineers, and technical staff

https://lz.lbl.gov/ @lzdarkmatter

- Black Hills State University
- Brandeis University
- Brookhaven National Laboratory
- Brown University
- Center for Underground Physics
- Edinburgh University
- Fermi National Accelerator Lab.
- Imperial College London
- Lawrence Berkeley National Lab.
- Lawrence Livermore National Lab.
- LIP Coimbra
- Northwestern University
- Pennsylvania State University
- Royal Holloway University of London
- SLAC National Accelerator Lab.
- South Dakota School of Mines & Tech
- South Dakota Science & Technology Authority
- STFC Rutherford Appleton Lab.
- Texas A&M University
- University of Albany, SUNY
- University of Alabama
- University of Bristol
- University College London
- University of California Berkeley
- University of California Davis
- University of California Los Angeles
- University of California Santa Barbara
- University of Liverpool
- University of Maryland
- University of Massachusetts, Amherst
- University of Michigan
- University of Oxford
- University of Rochester
- University of Sheffield
- University of Wisconsin, Madison

US UK Portugal Korea



Thank you!

Thanks to our sponsors and 35 participating institutions!

U.S. Department of Energy
Office of Science

Science and Technology Facilities Council

Sanford Underground Research Facility
South Dakota Science and Technology Authority

FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA EDUCAÇÃO E CIÊNCIA

Institute for Basic Science