Applications of Machine Learning to Big Data in Order to Identify Problematic Gamma-ray Spectra

WOSMIP 2022

STOCKHOLM, SWEDEN

KELLY TRUAX[1],
HENRIETTA DULAI[1],
MILTON GARCES[1],
THEODORE BOWYER[2],
JUDAH FRIESE[2], AND
LORI METZ[2]

[1]UNIVERSITY OF HAWAII AT MANOA,
[2]PACIFIC NORTHWEST NATIONAL LABORATORY

Background

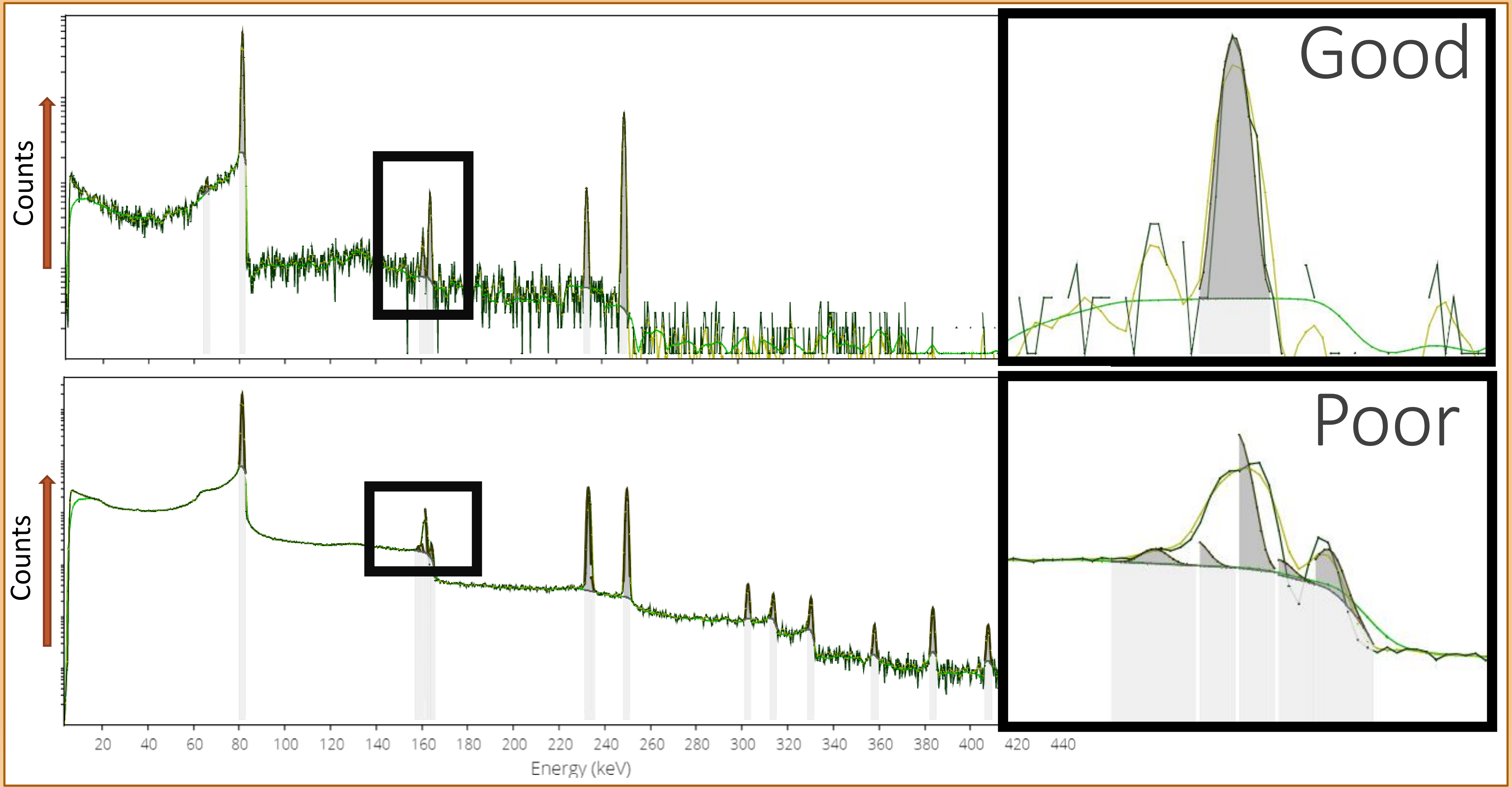
- Consortium for Monitoring, Technology, and Verification (MTV)
    - Signals and Source Terms for Nuclear Nonproliferation – DOE & NNSA
- Fosters opportunities for collaboration between Universities and National Laboratories – forum for ideas and discussion
- This project is an opportunity for a UH student to collaborate with PNNL on studying monitoring data sets – taking advantage of atmospheric radioisotope monitoring efforts by PNNL (STAX)
- Challenge: STAX has high data volume and multiple factors affect spectrum interpretation
- Goal: identify problems in detection and interpretation of spectra/peak fitting

STAX Data and Project Goals

- Xenon isotopes are used for monitoring and tracing atmospheric transport of radionuclides produced from industrial, power, and military sources.
- Reliable peak and activity (Xe activity ratio) analysis of spectra produced from continuous monitoring is needed.
- **Objective** - develop a method aimed at identifying spectra where automated peak fitting software (Genie 2000, AutoSaint) may fail to accurately report activities due to poor peak fitting.
- Identified spectra can then be re-analyzed manually and software corrections made if needed.
- Develop a method that integrates data from Genie 2000 and AutoSaint and identifies patterns of problems
  - Identify parameters best suited for flagging instances where poor fitting may result.
  - Flag spectra that need to be re-evaluated

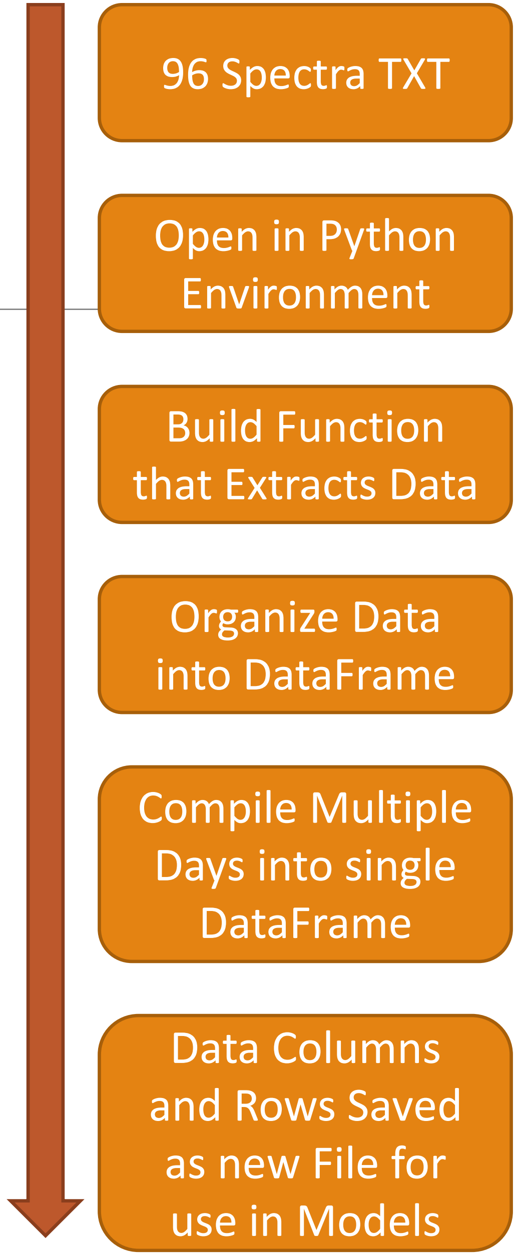
Isotopes of interest	γ energy [keV]
Xe-131m	164
Xe-133	81
Xe-133m	233
Xe-135	250
Xe-135m	527

Example STAX Gamma-ray Spectra with Software Fitting Xenon Peaks



Analytical Approach

- Data stored as individual 15-minute spectra and as daily 96 sample txt files
- Extract information from 96 sample files for both Genie 2000 and AutoSaint spectrum analysis
- Features interest:
  - Spectrum information: Time, Xe isotopes identified
  - MDA, MDC, and MDR
  - Peak information: Centroid, Counts, Uncertainty, FWHM
- Initially focus was given to peak specific data (Centroid, Counts, Uncertainty, FWHM)
- Data set for month of January 2022 was used for modeling tests – presented in Results.
- Data subjected to supervised and unsupervised model analysis to identify outliers in features.



96 Spectra TXT

Open in Python Environment

Build Function that Extracts Data

Organize Data into DataFrame

Compile Multiple Days into single DataFrame

Data Columns and Rows Saved as new File for use in Models

Supervised ML

- Data is labeled, sometimes as good/poor
- Learns from training data – predict unforeseen data
- Can be more time consuming with constant adjustment to the model to improve results
- Good for processing data output from previous experience
- Helps optimize performance from experience
- Have an idea of what to expect

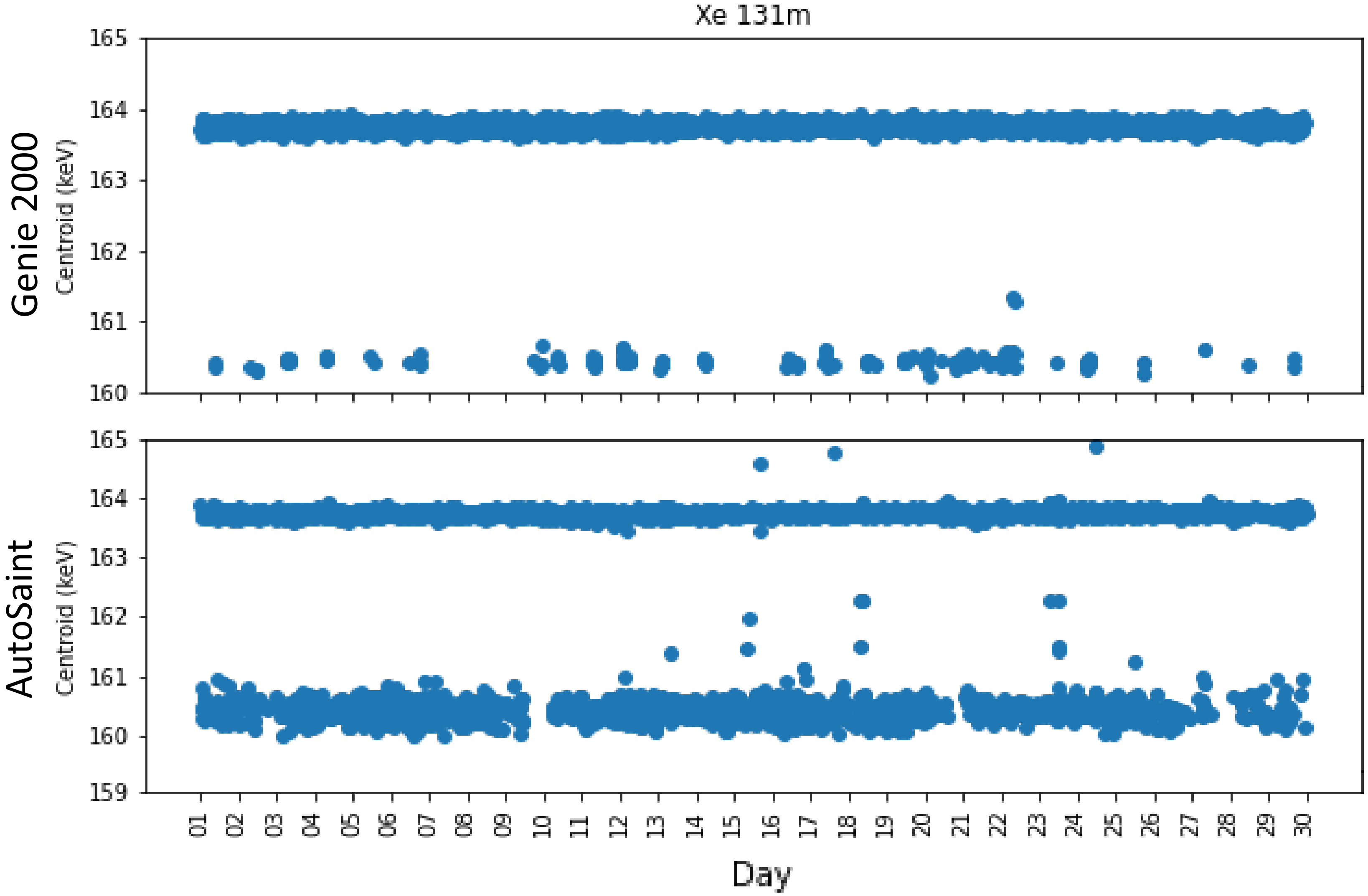
- User controls the inputs
- Can introduce unintended bias
- Example: Classification– yes/no

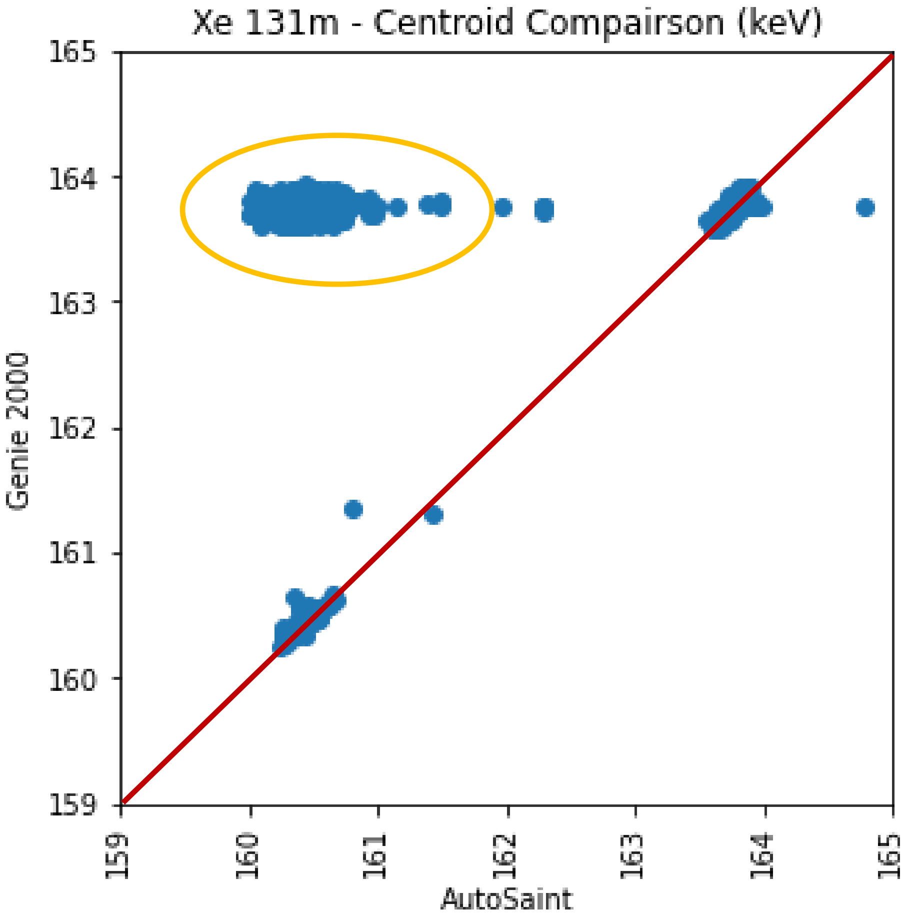
Unsupervised ML

- Technique where the model works to discover information – unlabeled data
- More unpredictable – needs verification
- Can handle more complex processing tasks
- Good for finding unknown patterns in data
- Find features to categorize
- Can be done in real time
- Better for anomaly/outlier detection

- No limit on inputs
- Uses data to make adjustments
- Example: Clustering – groups (how similar)

Supervised Model – Binning data reveals inconsistencies





Emerging Patterns

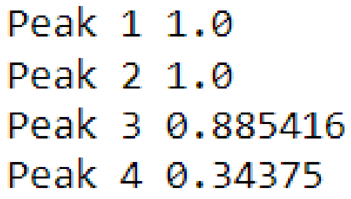
- Data binned into ROI based on energy representing isotopes of radioxenon
- Comparison of peak fitting with 2 methods within bins – found variation within peak reporting
- Multi-peak occurrences within these bins became a focus for identifying poor peak fitting using simple decision trees
- Both Software (Genie 2000 and AutoSaint), perform well when 2 peaks are present for Xe-131m
- Closer inspection of model results show poor fitting 100% of the time when 3 peaks present within a bin
- However, all of this is supervised approach prone to bias

Multi-Peak Occurrences (2880 Spectra)				
Isotope	**Software**	**Total #**	**Good Fit**	**Poor Fit**
Xe-131m	**Genie 200 2 peaks**	140	99%	1%
Xe-131m	**AutoSaint 2 peaks**	1636	98%	2%
Xe-131m	**AutoSaint 3 peaks**	11	0%	100%
Xe-135	**AutoSaint 2 peaks**	125	78% (245-246)	22% (247)
Xe-135	**AutoSaint 3 peaks**	18	0%	100%

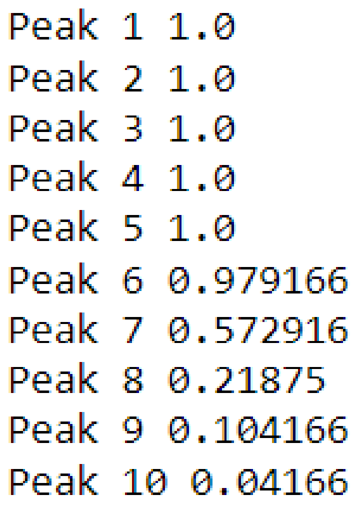
Unsupervised Model

- Unsupervised machine learning able to process larger volumes of data - increased data does not guarantee improved model performance
- Principal component analysis (PCA) was used to determine what in the bulk data is most likely to impact analysis
- Data sets from both Software's show varied reporting concerning the number of peaks. Only peaks reported >50% of the time should be included so as not to introduce bias
- 7 ROI bins were included for AutoSaint analysis
  - Peak Centroid, Peak Counts, Peak FWHM were included
  - 21 total features

Reported	Genie 2000	AutoSaint
Min # of Peaks	1	3
Max # of Peaks	4	10
Optimal # of Peaks for PCA	2-3 3	5-7 6

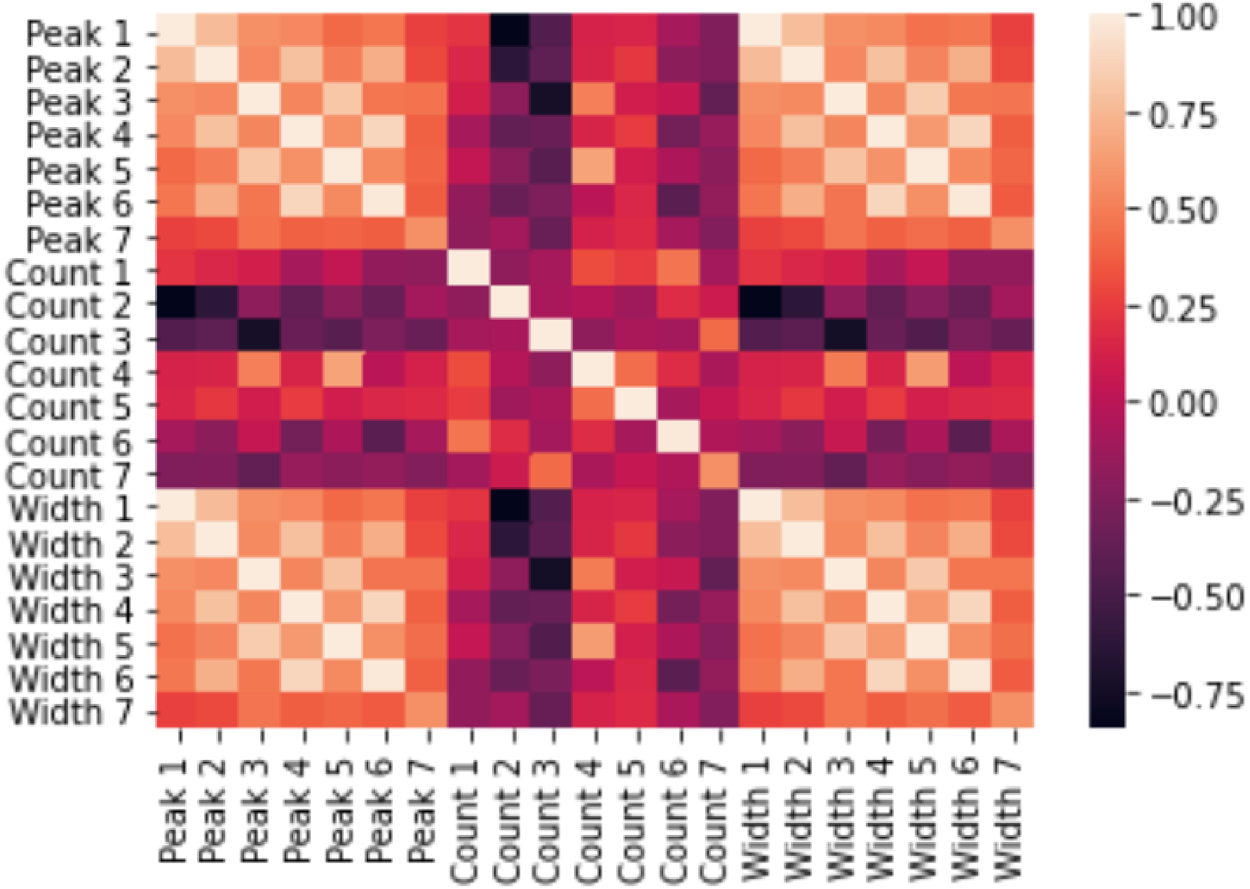


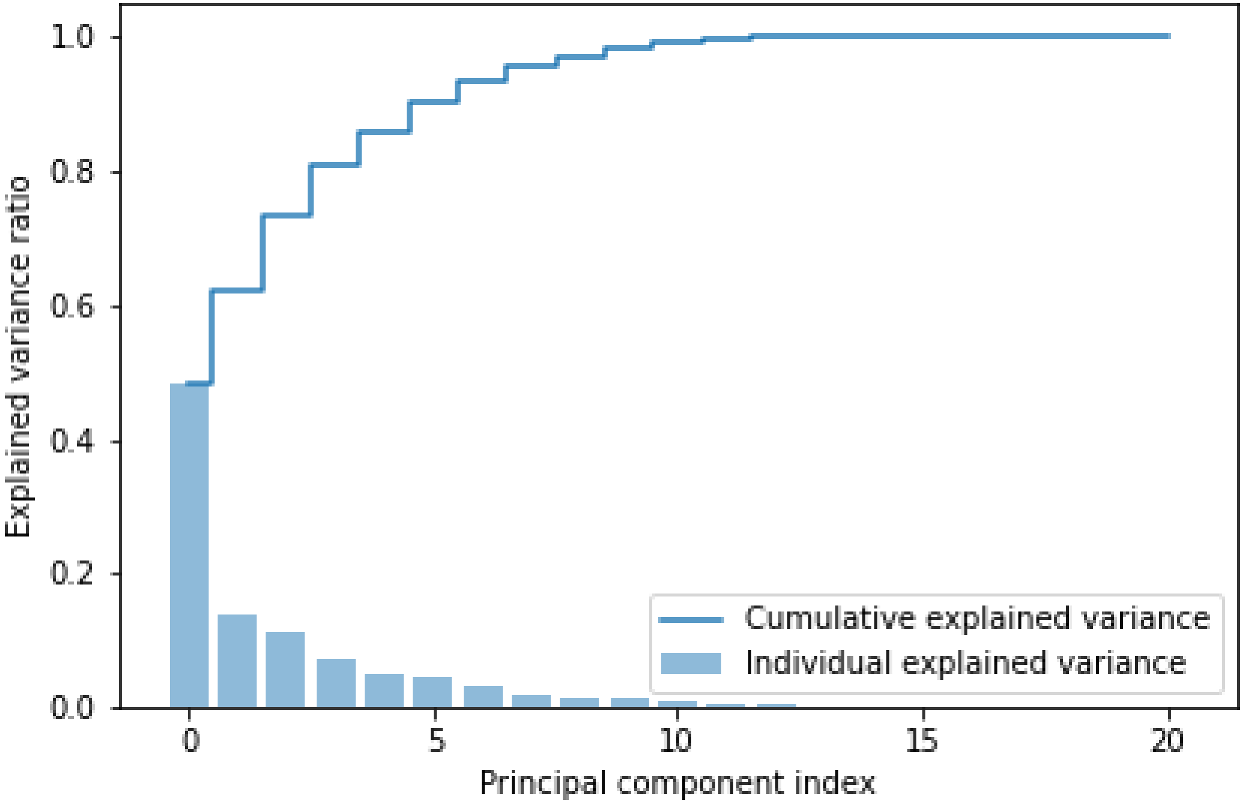
```
Peak 1 1.0
Peak 2 1.0
Peak 3 0.885416
Peak 4 0.34375
```



```
Peak 1 1.0
Peak 2 1.0
Peak 3 1.0
Peak 4 1.0
Peak 5 1.0
Peak 6 0.979166
Peak 7 0.572916
Peak 8 0.21875
Peak 9 0.104166
Peak 10 0.04166
```

Covariance Matrix and Principal Component Analysis

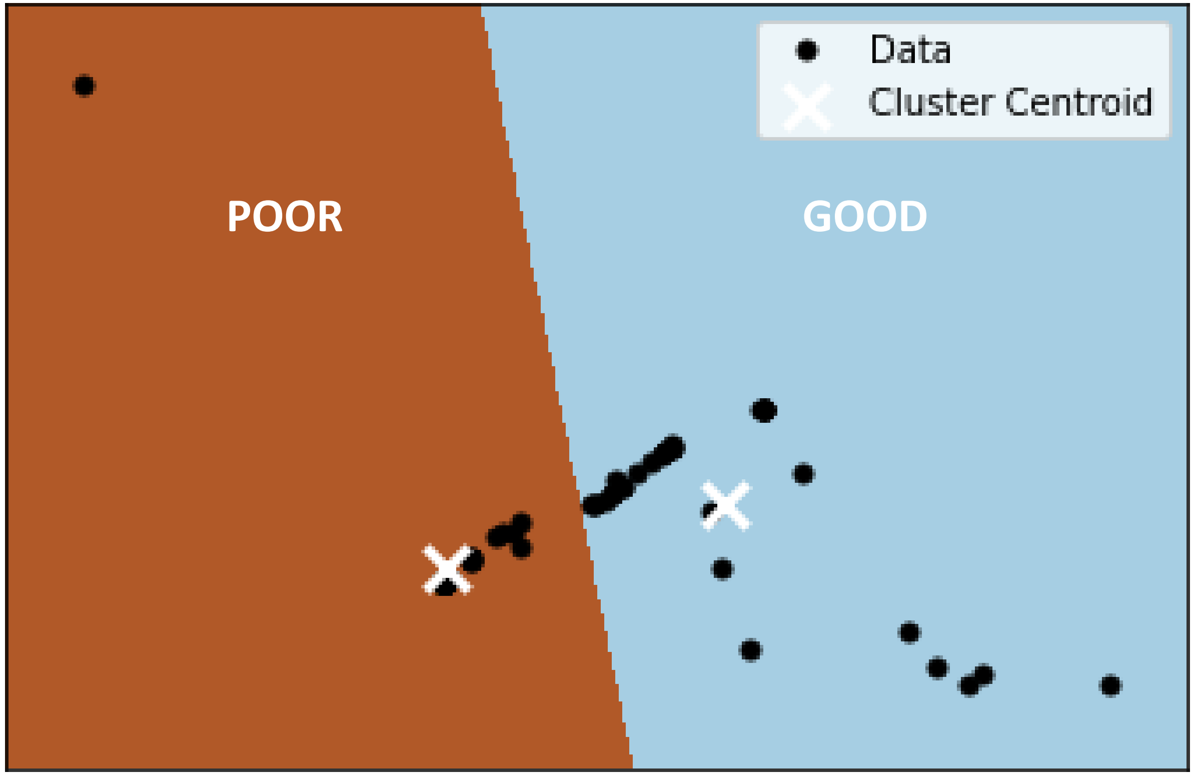




K-means NN Model

- Unlike a supervised model, the simple k-means is not told which data points are good or poor fits
- Clusters partition similar data into groups and separate from data farther apart. Similarity is determined by the distance between two points.
- Of 96 spectra for Jan. 15th, 2022
  - 16 spectra were identified by user as problematic
  - 13 identified by KNN as falling into POOR cluster
  - ~80% precision of basic model
- Success using only 10 principal components
  - First 6 peaks, first 4 counts
- Ran into problems processing full month of data due to memory demands

K-means clustering on the STAX dataset - 01/15/22

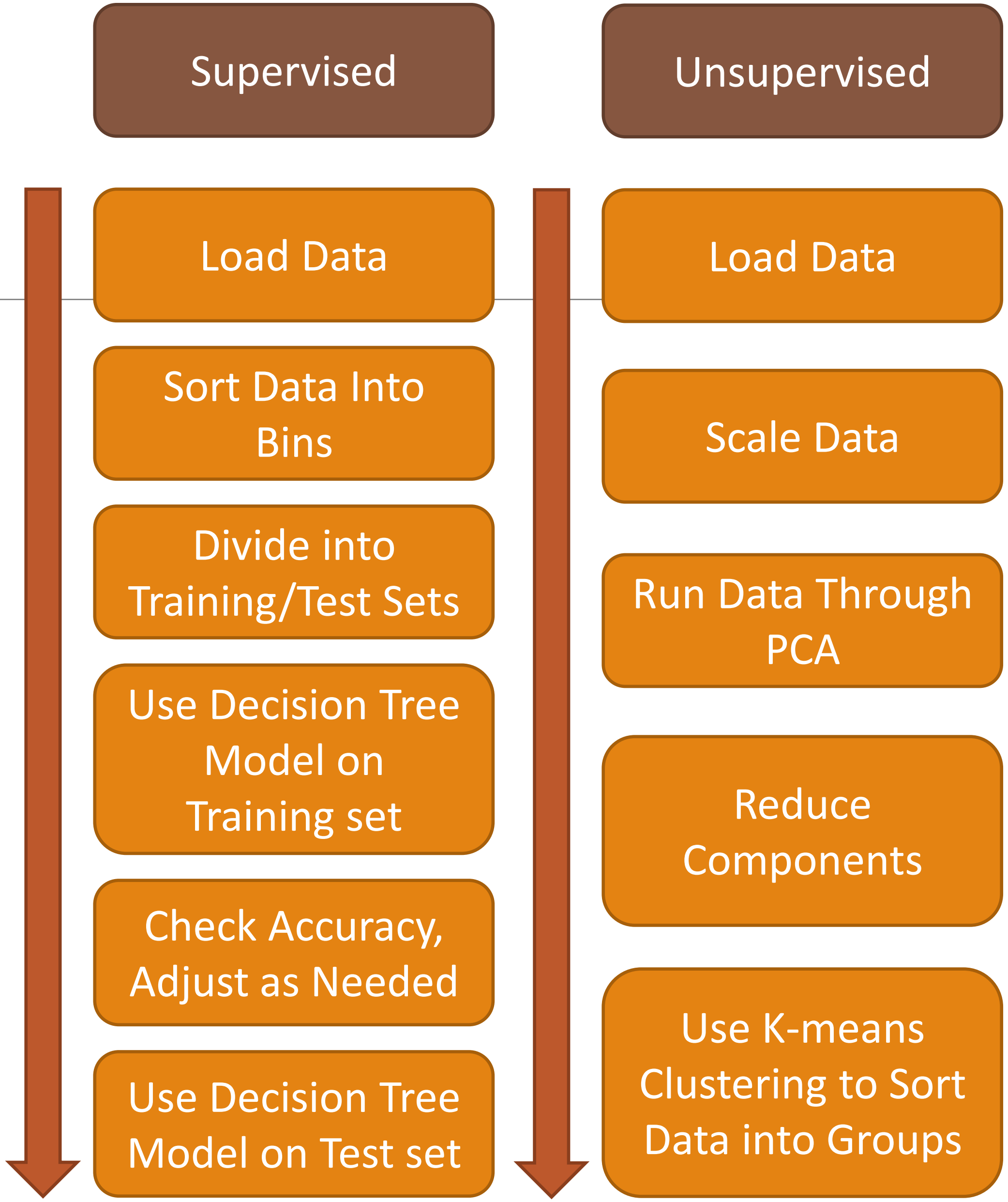


Summary and future work

- Initial observations show Genie 2000 is more consistent in peak fitting, while AutoSaint is more sensitive to the presence of multiple peaks.
- Further improvement of supervised and unsupervised approaches to improve accuracy. Can use supervised to check unsupervised results for consistency.
- Work to improve accuracy and test on more recent data to ensure reproducibility of results
- Examine data further to see if patterns emerge correlated to periods of high vs low count rates
- Success with unsupervised model allows for further expansion of data to explore potential impact of MDA, MDC, and MDR. Work to compress multiple days for memory use reduction may be needed using Neural Networks.
- Apply models to more recent data and record potential variation over longer period of time.
- Explore next step to implement real-time testing.

Conclusions

- Genie 2000 and AutoSaint correctly identify the presence of radioxenon isotopes within >95% of gamma spectra – need to be able to identify the 5% problematic ones
- Multi-peak occurrences within narrow energy (keV) ranges appear to be problematic as confirmed using supervised machine learning models.
  - Limited to predicting future data using previous cases
  - Can be hindered by user biases.
- Unsupervised machine learning was employed to handle the bulk data
  - Deduced that 6-7 ROI energy bins are optimal for analysis of AutoSaint files
  - Recognized patterns where peak counts could correlate to problematic fitting
  - Useful in identifying future peak data using k-mean cluster regions, PCA allows for quick scanning



Supervised

Load Data

Sort Data Into Bins

Divide into Training/Test Sets

Use Decision Tree Model on Training set

Check Accuracy, Adjust as Needed

Use Decision Tree Model on Test set

Unsupervised

Load Data

Scale Data

Run Data Through PCA

Reduce Components

Use K-means Clustering to Sort Data into Groups

Acknowledgements

Pacific Northwest National Laboratory - organization of WOSMIP 2022 and guidance on the work presented.

To the other Sponsors of the conference: the Swedish Defense Research Agency (FOI), the Comprehensive Nuclear-Test-Ban Treaty Organization (CTBTO), and the Information Science and Technology Institute (ISTI).

Finally, to NNSA, and DOE. This work was funded by the Consortium for Monitoring, Technology, and Verification under Department of Energy National Nuclear Security Administration award number DE-NA0003920