SAND2021-8725C

LDRD
LABORATORY DIRECTED RESEARCH & DEVELOPMENT
WHERE INNOVATION BEGINS

Automated algorithms for predicting trends and identifying subpopulations in neutron generator (NG) production data

R. Multari, C. Co, R.M. Ferrizz, L.M. Miller, E. G. Bujewski, J. S. Howard, M. Nelson , M. Dailey, S. Marklin, D. Kalinich, and J. Ray

Sandia National Laboratories

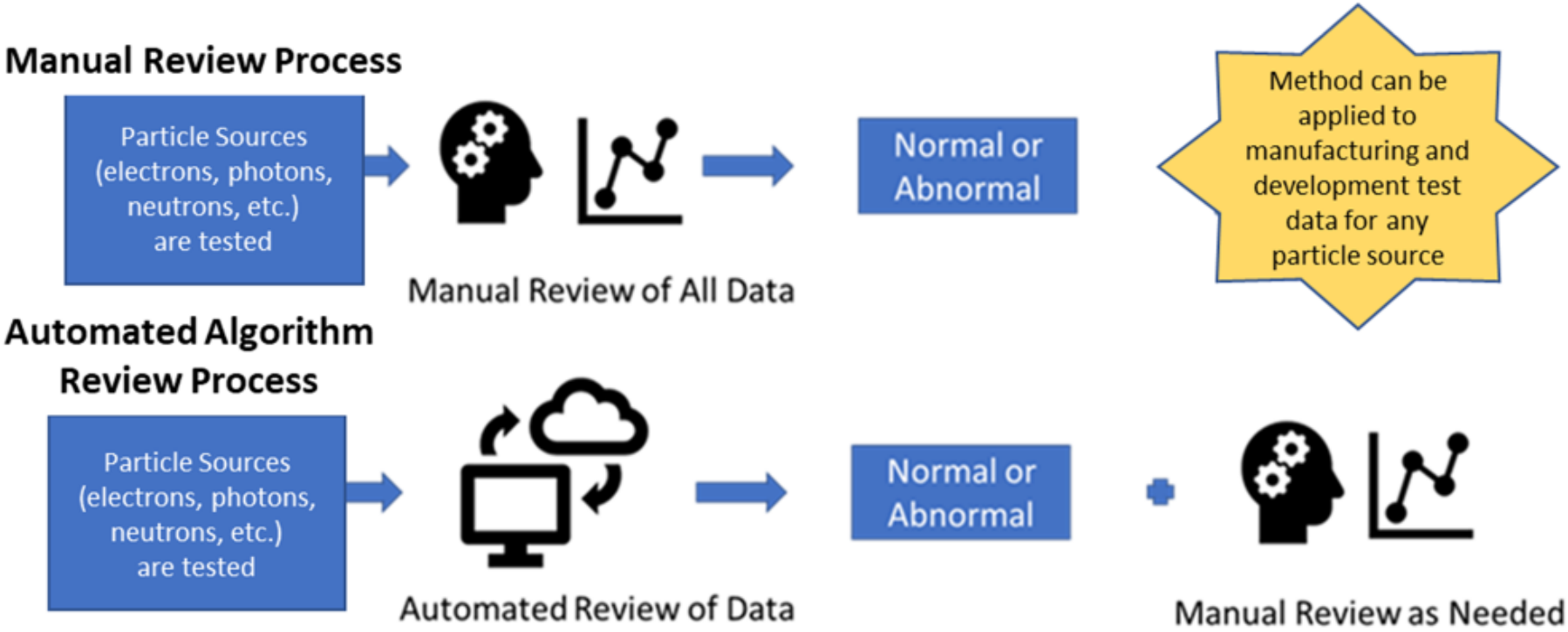
MLDL Workshop July 2021

To be covered

- Project background
- Applicability
- Development strategy
- Detection objectives
- Modeling data
- Methods
- PCA - Classifier modeling theory & results
- PLSDA modeling theory & results
- Ensemble algorithm & results
- Summary
- Acknowledgements

Project background

- Manual review of production test data is often required for the manufacturing of particle sources (neutron, electron, ion, etc.).
- The goal of this work is to create automated algorithms for the rapid, real time analysis of neutron generator test data.



LDRD - 3 year project (in year 3)

Year 1:
Demonstrate Feasibility Multivariate & Wavelet modeling

Year 2:
Investigate Capability Multivariate & Wavelet modeling

Completed Year 1 & Year 2

Year 2 results presented at MLDL 2020

Year 3:
Compare Multivariate and Wavelet modeling performance for different failures modes and create an ensemble algorithm using best methods

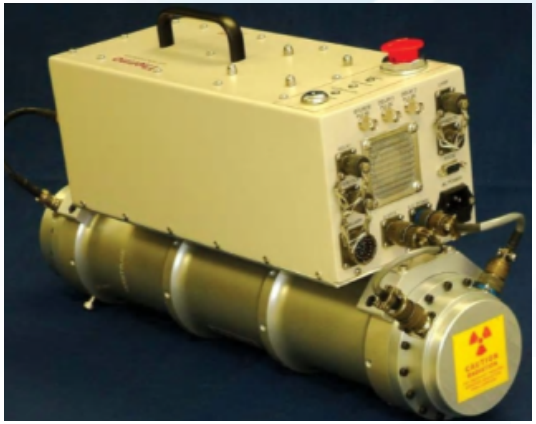
In Progress

Year 3 findings
Multivariate Modeling outperformed wavelet modeling

Ensemble algorithms created by combining two multivariate modeling methods

Applicability

- The methods being used for development can be applied to the development of automated analysis for other types of particle sources.



Neutron sources

Photon sources

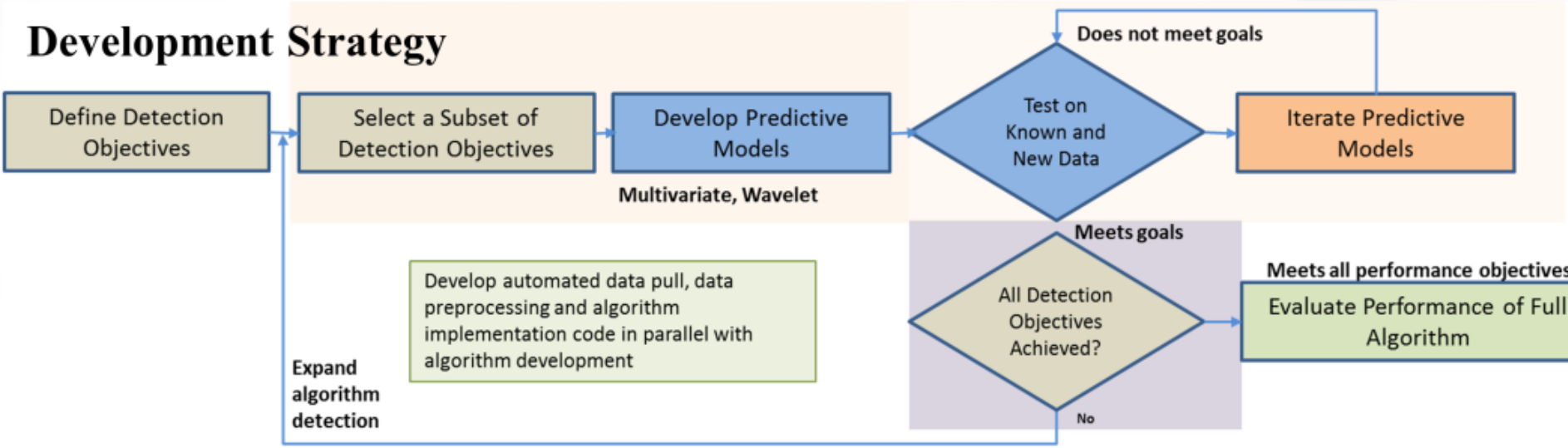
X-ray sources

SNL Patent Filed

R. Multari, J. Ray, L. Miller, P. Cummings, R. Ferrizz, L. Walla, N. Patel, S. Martin, and C. Co, United States Patent Application 17/172,353: **Systems and Methods for Screening Particle Source Manufacturing and Development Test Data**

Development strategy

- Select a set of known failure modes
- Create predictive models to identify abnormal test data & unusual behavior
- Using models with good predictive performance, construct an algorithm to successfully screen normal from abnormal data and identify failure modes
- Iterate process until all failure modes have been included



Data detection goals

- Normal data
- Abnormal data
  - Catastrophic fail type 1 – CF1
  - Catastrophic fail type 2 – CF2
  - Catastrophic fail type 3 – CF3
  - Passing but abnormal type 1 – PA1
  - Passing but abnormal type 2 – PA2

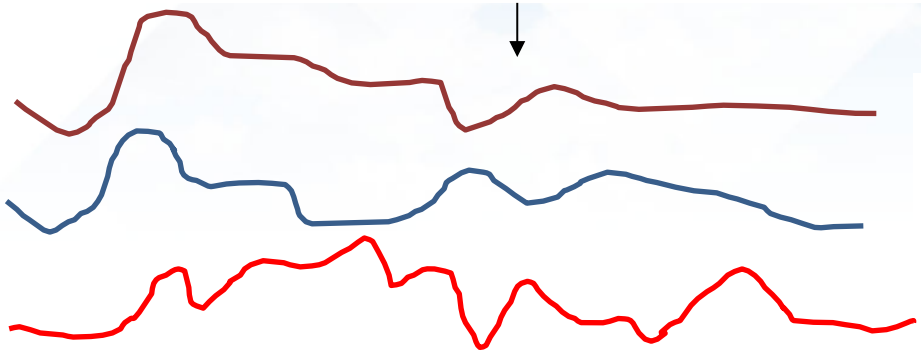
Sandia National Laboratories

Multivariate modeling data

Modeling data for a unit consists of all data collected during production testing of the unit appended together to make a single 1Xn array

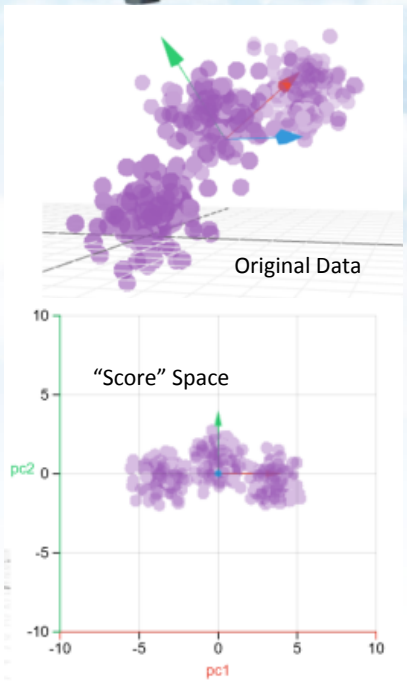
Voltages Currents Fluxes Etc.

Fingerprint created from all production test data



Multivariate modeling theory

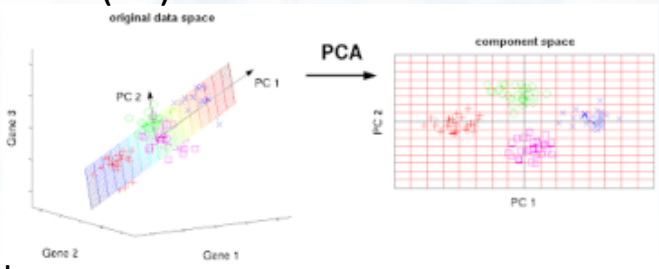
- Assumes a relationship exists between a set of measured variables and the properties of interest
- Observation = Structure + Noise
  - Variables X (set of observations)
  - Response Y = F(X) (set of possible responses)
- Finds the structure in the data representing the correlation between F(X) & X
- Goal of the modeling is to extract the structure in the data that correlates to the observed responses while minimizing noise
- Analysis is accomplished through successive transformations in which the data is projected onto axes or "Principal Components" (PC's) representing the direction of maximum variation of the data
- Each PC is orthogonal to the other PC's and centered on the mean of the data and is aligned to the direction of the maximum variation of the data
- With each successive transformation to a new PC, more of the variance in the data is explained and a smaller portion of the variance remains unexplained

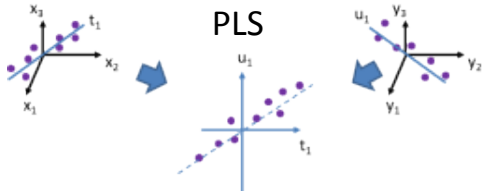


Projection of original data onto PC1 & PC2

Multivariate modeling methods used

- Two predictive multivariate modeling methods used
  - Principal Component Analysis (PCA) - based classifiers
  - Partial Least Square Discriminate Analysis (PLSDA)
- PCA-based classifiers
  - Based on combining output waveforms into orthogonal bases (principal components)
    - A quick way of doing dimensionality reduction (Scree plots with PCs)
  - Equally easy to use with a Naïve-Bayes (NB) or Random Forest (RF)
  - Cluster visualization (via t-SNE*)
- PLSDA predictive models
  - Very sensitive to differences in data groups
    - Generated using commercial software (The UnscramblerX)
    - Analysis algorithm constructed using combinations of models
  - Very good for group identification
    - Hones in on features most different among the groups



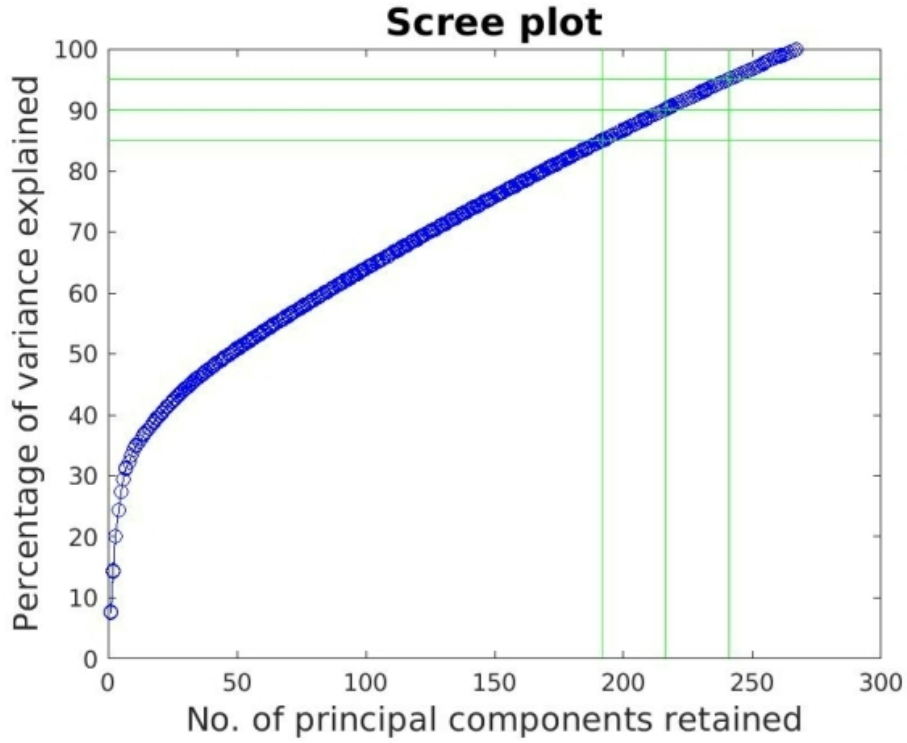


*t-distributed stochastic neighbor embedding

Based on creating probability distributions such that similar high-dimension data has a higher probability and dissimilar data a lower probability

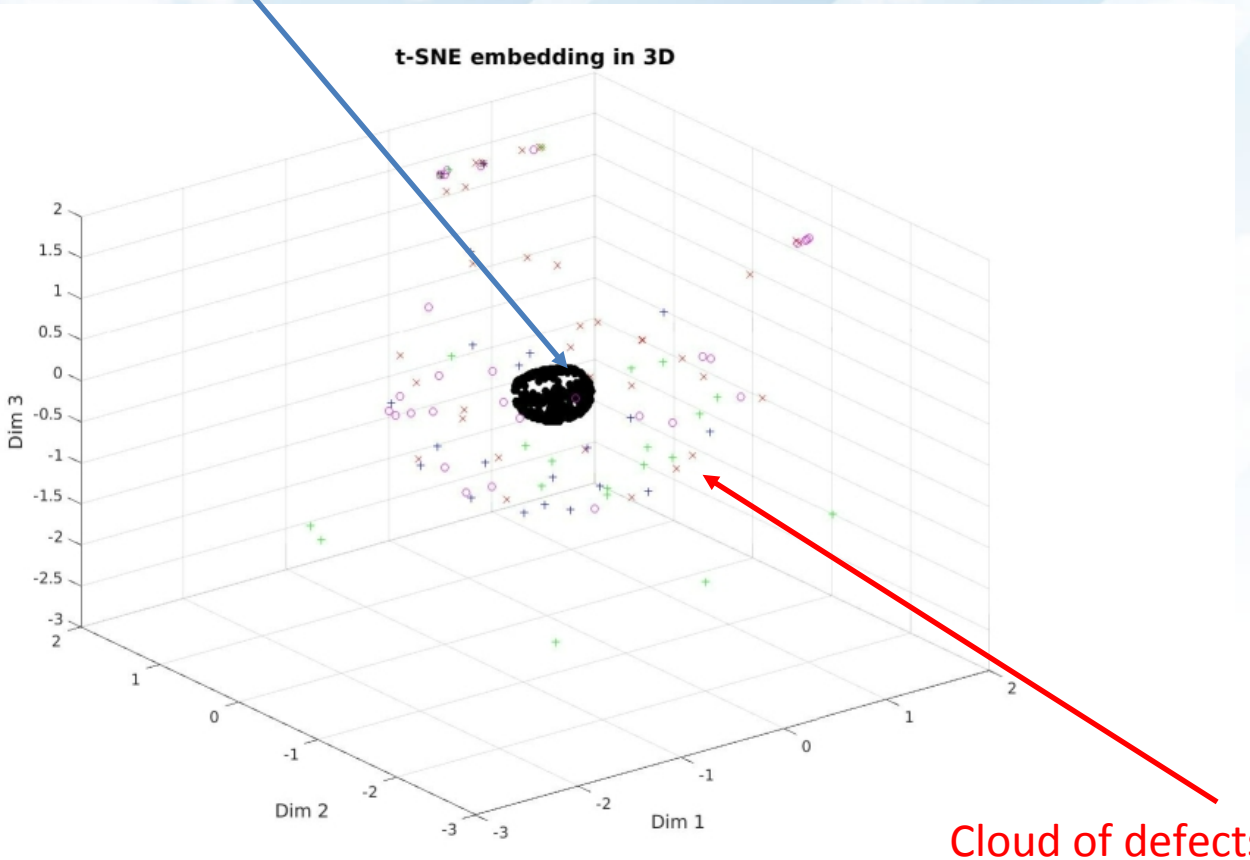
PCA-based classifier theory

- Concatenate all waveform test data
  - “fingerprint”; vector, 80,000 long
- PCA the fingerprints
  - Choose 216 principal components (90% explanation)
- Dimensionality reduction
  - From 80,000 to 216
- Use in Random Forest and Naïve-Bayes



PCA-based classifiers

Cluster of 'Normal' Devices – tight cluster



- Used to determine normal vs. abnormal test data (waveforms)
- Are the normal & abnormal test data separated in 216-dim PCA-space?
  - t-SNE plot
- If yes, classifier can be used to separate the groups

Cloud of defects; no cluster

PCA classifier results: NB vs RF

- Make a balanced training dataset
  - 138 Normal, 138 Abnormal
- Train a Naïve Bayes classifier
  - Repeated random sub-sampling
  - Misclassification: 0 (after 20 rounds)
- Train a Random Forest classifier
  - Misclassification rate: 1.09 % (mean over 29 rounds)
  - Misclassifies certain Abnormal as Normal (False negatives)
- Conclusion: Use NB classifier
  - Better performance

	Predicted Normal	Predicted Abnormal
GT Normal	100%	0%
GT Defective	0%	100%

NB classifier

	Predicted Normal	Predicted Abnormal
GT Normal	100%	0%
GT Defective	2.18%	97.82%

RF classifier

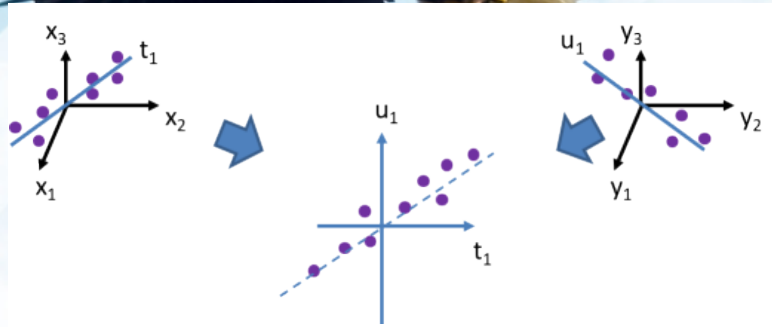
LDRD LABORATORY DIRECTED RESEARCH & DEVELOPMENT

Sandia National Laboratories

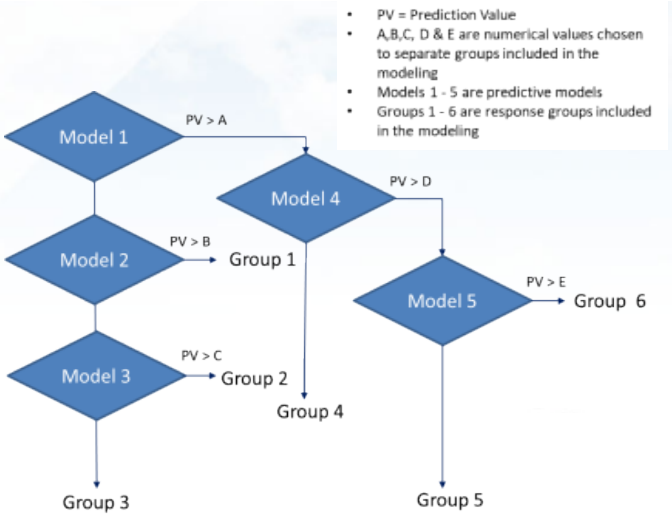
Partial Least Square Regression (PLSR)

- Data is arranged in a 1Xn matrix for modeling
- PC's are calculated by modeling both the X and Y matrices (variables and responses) simultaneously using known data
  - Uses PCA on the variables (X^TY)
  - Uses PCA on the responses (Y)
  - Creates a transformation designed to maximize the covariance between X & Y
- Each interactively calculated PC has a characteristic linear equation for the relationship of the response to the variables :

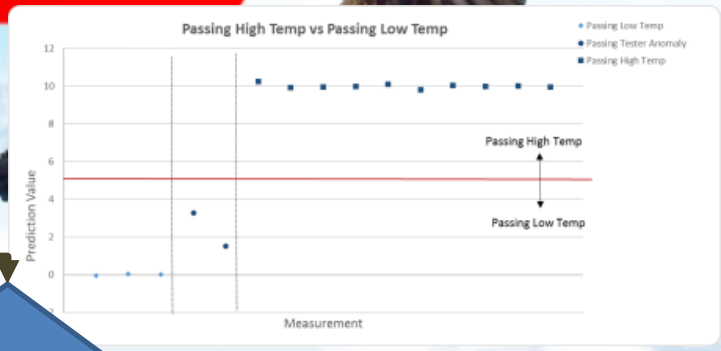
  $Y = b0 + b1x1 + b2x2 + b3x3 + \ldots$
  - The loadings indicate the contribution of each variable to the PC calculation
- Using an optimal number of PC's, a "Prediction Value" (PV) is calculated by the PLS prediction model that indicates how well matched new input data is to one of the response groups in the modeling
- Multiple models can be combined to create a programmed flow to differentiate new data based on PV's for input data



Y data structure influences the decomposition of the structure in the data



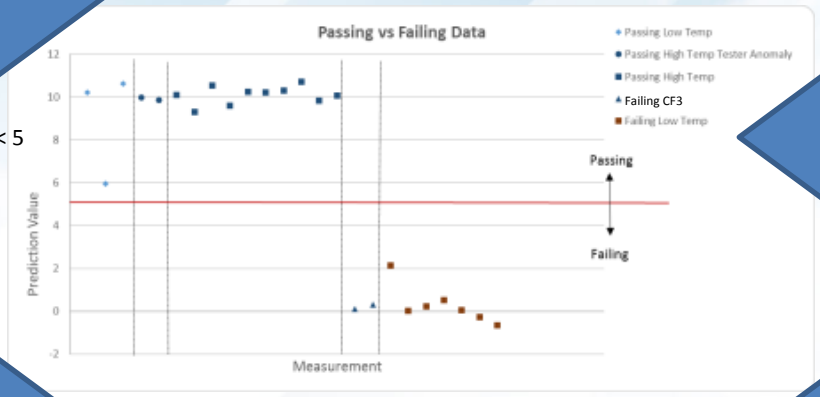
Multivariate modeling algorithm



Passing or Failing?

PV > 5

PV < 5



High or Low Temp?

PV > 5

Passing High Temp

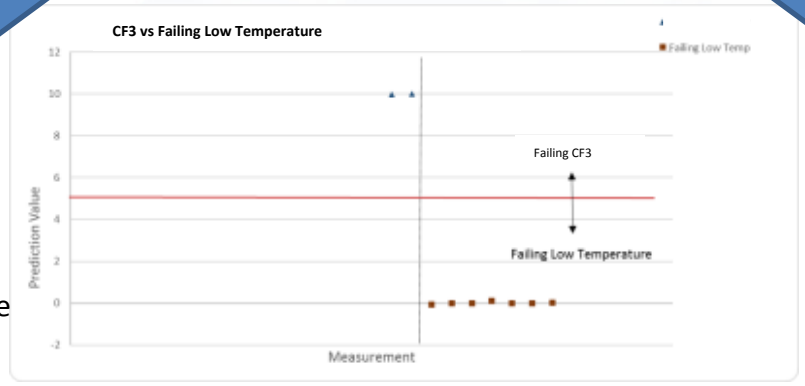
PV < 5

CF3?

PV > 5

CF3

PV < 5



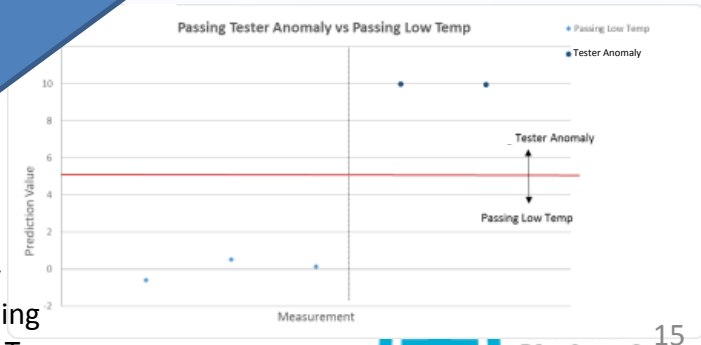
Low Temperature Fail

Tester Anomaly ?

PV > 5

Tester Anomaly

PV < 5

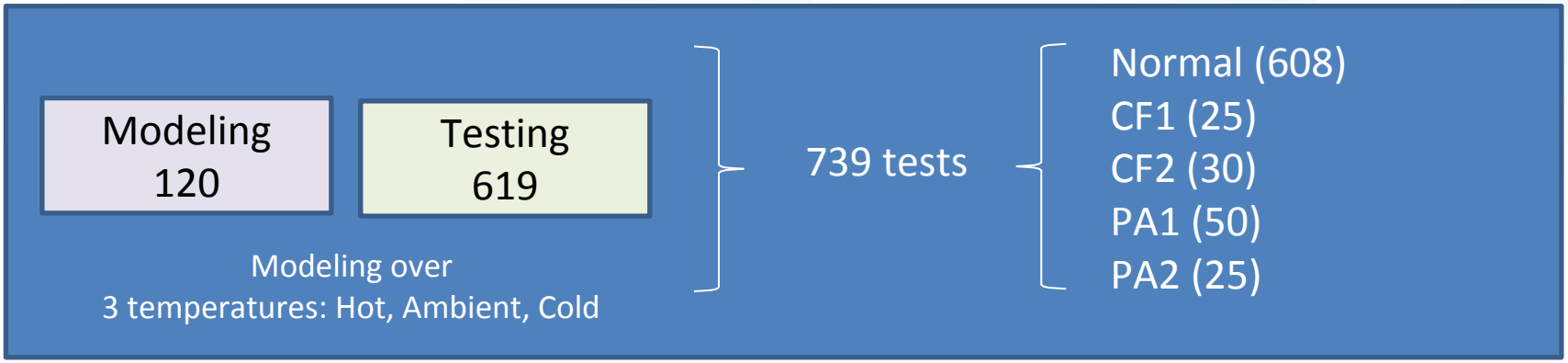


Passing Low Temp

National Laboratories

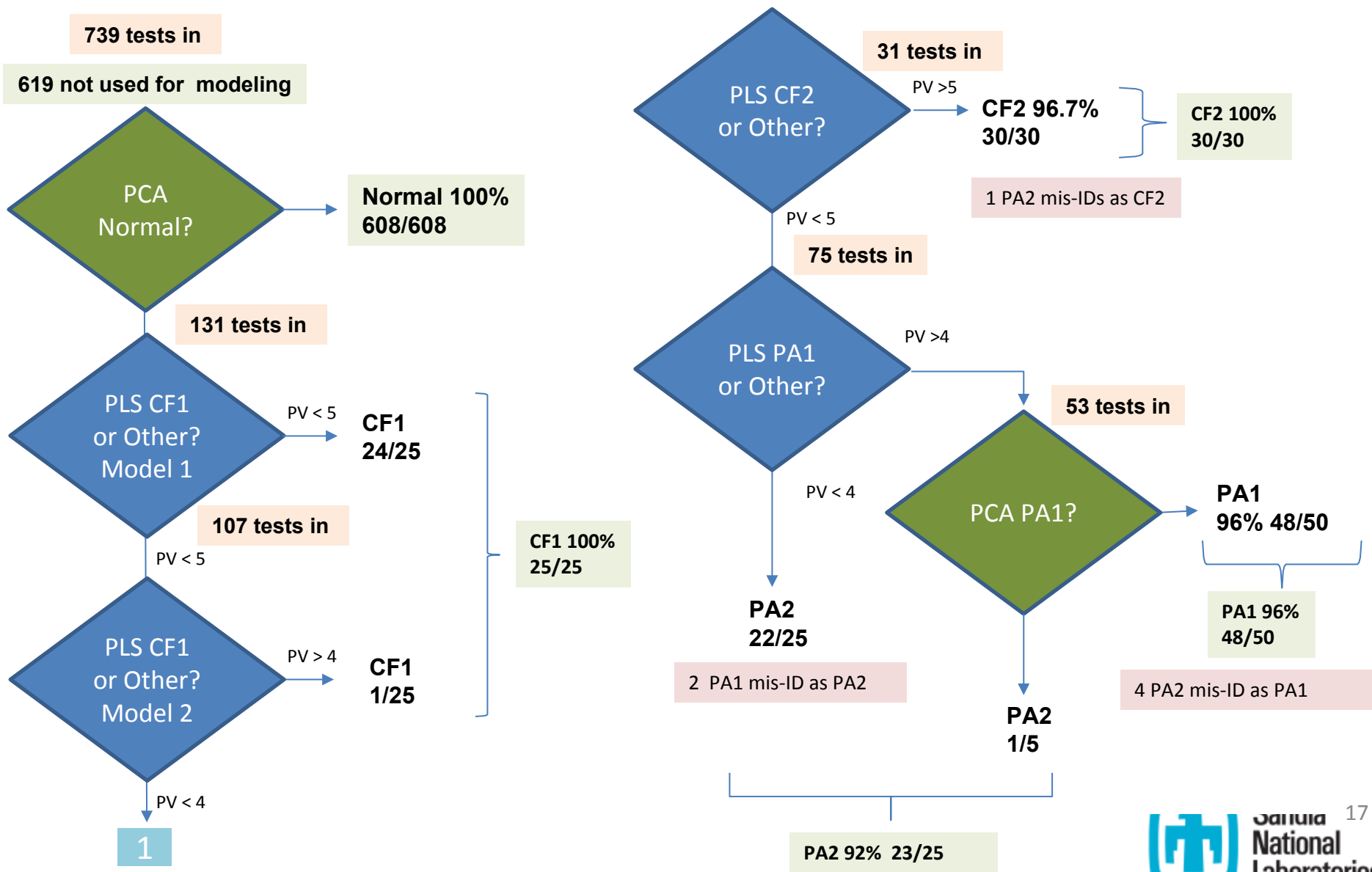
Ensemble algorithm

- Evaluated different PLSDA and PCA models
- Found PCA Classifier modeling out-performed PLSDA modeling for the separation of normal from abnormal test data and had to be used in combination with PLSDA modeling for the separation of PA1 from PA2 test data
- Ensemble algorithm was constructed from best performing PCA and PLS models



16

Ensemble algorithm results: 99.1% Correct ID



Summary

- Particle source manufacturing test data can be analyzed for performance without analyzing for specific parameters
- It is possible to design an ensemble algorithm using multiple multivariate modeling methods to correctly differentiate all data groups
  - Normal, abnormal, type of abnormality

Acknowledgements

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration. Funding for this research was provided by National Nuclear Security Administration (NNSA) Laboratory Directed Research and Development (LDRD).

Sandia National Laboratories